30(3):11-24; septiembre, 1999

Lenguaje natural e Indización automatizada

José A. Moreiro González
Eva Mª Méndez Rodríguez

Se plantea una aproximación teórica a la indización automática, en donde se pone de relieve el papel que ha desempeñado el lenguaje natural, no controlado, en su evolución, y se señalan sobre todo las últimas tendencias en indización automatizada fundamentadas en bases de conocimiento.

Introducción

Vivimos en un mundo esencialmente lingüístico en el que las cosas son lenguaje y el lenguaje es una cosa. La cultura, la producción científica y en, definitiva, el conocimiento que aporta al ser humano el dominio de la realidad, se conforma, se construye y difunde a través del lenguaje. El hombre piensa, lee, y escribe gracias al lenguaje (al lenguaje natural) de tal suerte que su código se erige como un potencial comunicativo.

En este contexto de la comunicación humana, la Documentación presenta una estructura lingüística [1] ya que el discurso sobre el que se emiten los datos se ejecuta en lenguaje natural, como *un aluvión de estructuras cognitivas en lenguaje natural.* Si bien es cierto que el lenguaje natural es aquel conjunto de signos y símbolos orales y escritos por medio de los cuales los seres humanos se comunican entre sí, dentro de este trabajo definiremos lenguaje natural como aquel conjunto de palabras utilizadas por un autor para expresar sus ideas en un documento.

Es evidente, pues, que existe una estrecha relación entre la Lingüística y la Gestión de la información, que podríamos explicar haciendo una extrapolación del concepto saussuriano de *signo lingüístico*, compuesto por significante (plano de la expresión, esto es, los grafemas que componen los términos de los documentos científico-técnicos) y significado (plano del contenido, o de la esencia semántica de los conceptos sobre los que se realiza el análisis de contenido en Documentación) (Fig. 1)

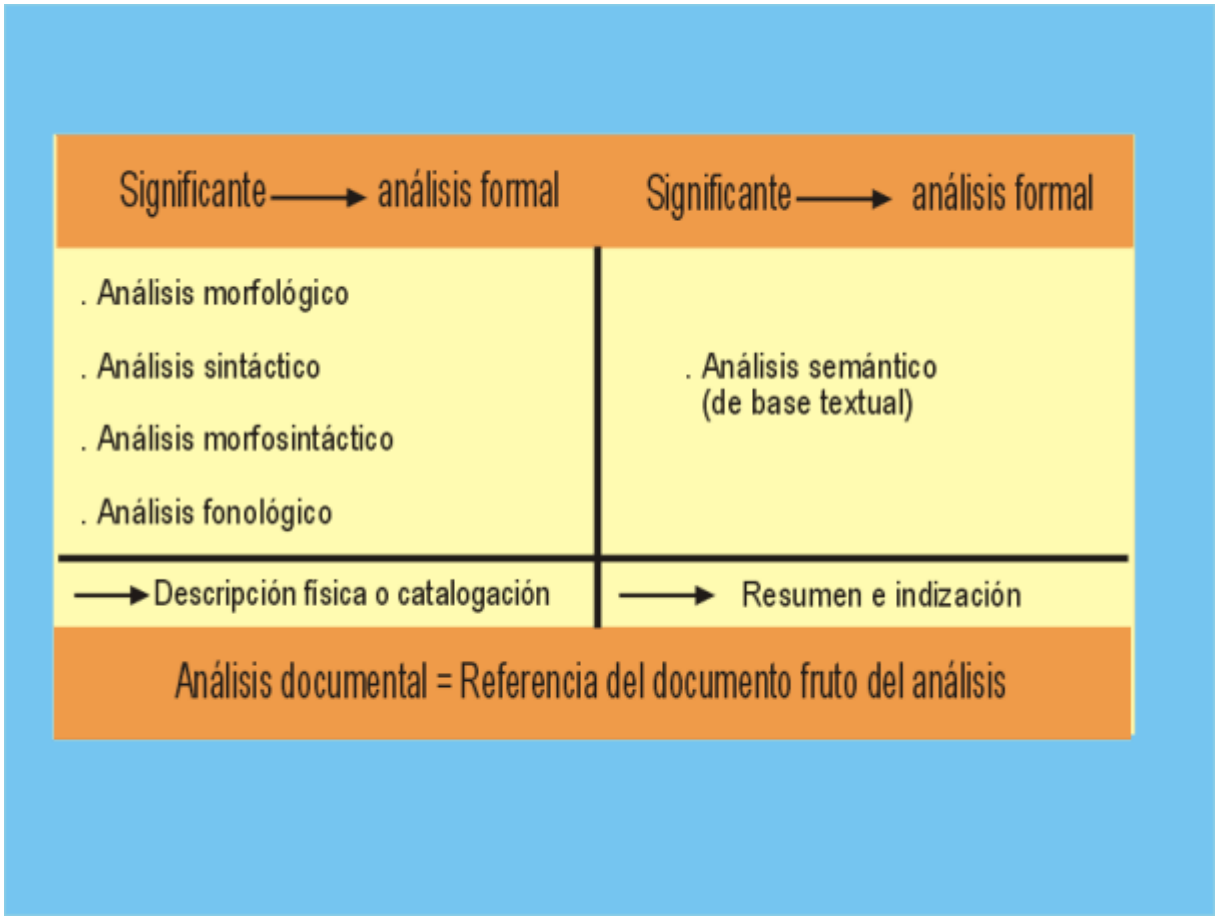

Fig. 1.

A pesar de la omnímoda implicación de la Lingüística en el Análisis Documental, en esta aproximación nos centraremos en la semántica, y concretamente en la semántica informática de cuyo desarrollo depende en gran medida la indización automatizada.

La comunicación científica se establece en lenguaje natural, un lenguaje que en su expresión escrita adolece de serias ambigüedades e imprecisiones derivadas precisamente de la falta de significado unívoco y preciso de las palabras que lo componen; presenta múltiples dificultades para el tratamiento de la información al estar compuesto por decenas de miles de palabras, y estar sujeto a diferentes accidentes léxico-semánticos (como la homonimia, polisemia, sinonimia, y figuras retóricas como anfibología, metáfora, símil, metonimia, anáfora, sinécdoque, etc.) que impiden la univocidad del signo lingüístico, y por ende, la comunicación exacta.

Pese a ello, hoy, el tratamiento y la recuperación de información en lenguaje natural es posible gracias a la intervención del ordenador. Cada vez son más abundantes los softwares documentales basados en el lenguaje natural que se destinan a interrogar bases textuales constituidas tanto en lenguaje cotidiano como en una terminología especializada. La trascendencia de estos programas para el tratamiento y la indización del lenguaje natural aumenta en el contexto en que nos encontramos: la explosión de la información textual posibilitada por ordenador, donde la edición electrónica a finales del siglo XX se ha convertido en un hecho a la vez que un problema para la recuperación de información. Por ello, a través de este trabajo, proponemos mostrar cómo ha evolucionado la indización automática en la gestión de las palabras desde los inicios en lenguajes absolutamente libres, hasta el momento presente determinado por la

regularización de las palabras en términos contrastados mediante tesauros y bases de conocimiento.

De la indización a la indización automatizada: justificación

La indización ha sido tradicionalmente uno de los temas más importantes de investigación en Documentación, ya que los índices han facilitado la recuperación de información tanto en los sistemas manuales tradicionales como en los nuevos sistemas informatizados. La indización *per se* está abocada a la recuperación de información. Con las oportunas salvedades históricas, podríamos decir que el concepto de recuperación de información es tan antiguo como el mundo escrito, y se magnifica su importancia cuando hablamos de un *mundo informativo digital*, en el que numerosas representaciones del conocimiento humano se hacen en formato electrónico.

La indización es uno de los procesos fundamentales del análisis de contenido, y son muchas las definiciones que se han dado pero todas ellas la definen como una técnica, la de caracterizar el contenido tanto del documento como de las consultas de los usuarios, reteniendo las ideas más representativas para vincularlas a unos términos de indización, bien extraídos del lenguaje natural empleado por los autores, o de un vocabulario controlado o lenguaje documental seleccionado *a priori*. Hoy en día es posible vincular el proceso de indización al lenguaje natural del documento gracias a los computadores; para hacerlo debemos discriminar la información aprovechando las estrategias utilizadas por los propios autores para presentar sus publicaciones, pues destacan la información esencial en títulos, resúmenes, y en los párrafos iniciales de las diferentes partes de los textos. También nos valemos de otras estrategias sintácticas y semánticas, como las que se derivan de la función que cumplen las palabras en las oraciones y del peso semántico que tienen las palabras en los textos. Si optamos por manejar el texto completo, sólo será posible una recuperación eficaz en aquellos lenguajes cuyos términos gocen de gran estabilidad. Tal sucede en los propios de las ciencias aplicadas y de la tecnología, donde la búsqueda se podría hacer en las mismas expresiones usadas por el autor.

Lo más frecuente es que el texto original y su *traducción* documental se den dentro de los dominios propios de las distintas áreas del saber. En este caso la amplitud de uso de los términos, de la expresión y del estilo que es propia del lenguaje natural, se ve limitada por las características fundamentales del discurso científico, lo que favorece la pertinencia de uso del lenguaje natural con fines documentales [2]:

- Recepción y emisión cualificada (competencia).
- Vocabulario especializado.
- Organización estructural útil a la ciencia.
- Modelado lógico-formal.
- Determinación más sistemática que el lenguaje común.

De igual forma que en la indización manual, el principio de indización automatizada es identificar un documento por un conjunto de palabras clave representativas de su contenido, que pertenezcan a un conjunto abierto de términos –indización libre–, o que pertenezcan a un conjunto cerrado y referenciado en una lista de autoridad o en un tesauro —indización controlada–. Así, pues, podemos definir la indización automatizada como el uso de máquinas para extraer o asignar términos de indización sin intervención humana, una vez se han establecido programas o normas relativas al

procedimiento.

Los factores que hacen posible pensar en el paso de una indización manual a una indización automatizada son, los siguientes:

- Alto coste de la indización humana (tiempo).
- Aumento exponencial de la información electrónica y la proliferación del *full-text.*
- La gestión electrónica de documentos (GED) y la informatización de los procesos documentales.
- Automatización de los procesos cognitivos y la investigación creciente y los avances en el Procesamiento del lenguaje natural (PLN) .

a) Alto coste de la indización humana en términos de tiempo es uno de los argumentos más sólidos que se ostentan para justificar el desarrollo de sistemas de indización automatizada.[1] Cómo explotar de manera pertinente con un coste y tiempo reducidos, el volumen siempre creciente de información textual, se ha convertido en un tema recurrente y obsesivo en todos los estudios de análisis documental de contenido, dando lugar a múltiples trabajos destinados a evaluar la coherencia y la pertinencia de indización automática frente a la humana.[2]

Otros autores [6, 7] encuentran la justificación de las investigaciones en indización automatizada, partiendo de la base que la indización humana es inadecuada para minimizar la subjetividad inherente a la indización, ya que el grado de consistencia alcanzado, depende no sólo del conocimiento de técnicas de abstracción conceptual, ni del conocimiento y manejo de lenguajes documentales, depende también del grado de conocimiento que el analista tenga sobre el tema que se trata, exigiéndole que esté siempre actualizado en esa materia. Es importante señalar también la inconsistencia entre los indizadores e incluso de un mismo indizador en distintos momentos anímicos, ya que la indización es algo subjetivo; el ser humano utiliza el lenguaje en función de múltiples condicionamientos, parcialidades y sesgos personales y culturales involuntarios.

Le exacerbación de *lo humano* como sinónimo de *lo racional* y lo perfecto es fruto del conservadurismo y de la fidelidad a la idea de ser humano, pero objetivamente desde el punto de vista de la indización o descripción característica del contenido de un documento, hay muchos casos de malos ejemplos en que la indización manual, es a todas luces, deficiente. Por tanto, todas estas argumentaciones nos han llevado a pensar que la indización automática es la formalización y/o automatización de la indización, con el objetivo de reducir la subjetividad del proceso, y el alto coste en tiempo de la indización manual.

b) El aumento exponencial de la información electrónica y la proliferación del *full-text*. En este sentido es interesante evocar la afirmación que hacía Jones en los años 80: “El valor de la indización automática se incrementará cuando la literatura de forma legible a máquina sea más importante que la producida por medios tradicionales. Entre tanto, el ordenador será de importante ayuda para el indizador en la elaboración de los índices, aliviándole de tareas rutinarias como la ordenación, clasificación e impresión. No obstante, por el momento, las acciones específicas de determinar lo que constituye la materia indizable del texto, y cómo se debe expresar, son funciones todavía de la inteligencia y creatividad humanas [8, p. 12].”

Esta afirmación que Jones hacía en 1986 como futurible, parece que es una situación del presente, no porque la literatura producida en forma legible por máquina sea más importante que la producción impresa, pero sí hay que tener en cuenta que la propia naturaleza de la información ha cambiado y cada vez más se presenta en formato electrónico. El crecimiento exponencial de cantidades de información producidas y/o reproducidas en redes internet e intranet es hoy ya una realidad; por ello parece inevitable que el valor de la indización automatizada se incremente y tienda a dominar con respecto a la indización tradicional humana.

El incremento de la ciencia y de la comunicación electrónica, crece de manera imparable; cada vez son más las bases de datos que se pueden consultar a texto completo, al mismo tiempo que la vida media de la información tiende a disminuir, todo ello contribuye a que no exista un paradigma unificado para la recuperación de información. La tarea de convertir en accesibles todas estas informaciones relevantes requiere una serie de actividades que componen el ciclo documental, entre las cuales, el análisis de contenido tiene un papel fundamental, con lo cual es lógico que las investigaciones en documentación busquen nuevas alternativas para optimizar la recuperación de información. Una de estas alternativas es la indización automatizada donde, acudiendo a otras disciplinas como la lingüística o la estadística, se pretende dar solución al problema de la caracterización del contenido documental, y con ello, de la recuperación de información.

c) La gestión electrónica de documentos (GED) y la informatización de los procesos documentales. Las organizaciones están asumiendo en la actualidad una tendencia incipiente de conversión de los archivos basados en papel a los sistemas de gestión electrónica de documentos (EDMS: *Electronic Data Management Systems)*. Esta tendencia supone una nueva filosofía en el tratamiento de la documentación, combinando la imagen con la información textual asociada a ella, que requiere una planificación exhaustiva, donde la indización de documentos digitales insta un proceso informatizado de comprensión e inferencia del contenido para su posterior integración y recuperación en los procesos.

La automatización de los procesos documentales –almacenamiento, recuperación y reproducción de los documentos– mediante herramientas y aplicaciones informáticas, está estrechamente ligada a la indización automatizada, ya que la mayoría de los sistemas GED incluyen un motor de indización y búsqueda para procesar el lenguaje natural y efectuar la recuperación por contenido.

d) La automatización de los procesos cognitivos y la investigación creciente y los avances en el procesamiento del lenguaje natural (PLN). Existen numerosas metáforas antropomórficas aplicadas a las máquinas en el sentido de que la eficacia en el procesamiento de la información es la característica esencial que comparten el ordenador y la mente humana.

La mente humana posee una eficacia cualitativa en sus procesos cognitivos (percepción, decisión, planificación y lenguaje). Existen distintas teorías que avalan que el lenguaje natural, lenguaje de comunicación humana, no es un lenguaje interno de pensamiento sino que es un lenguaje fruto del aprendizaje. De esta afirmación, podemos deducir que las máquinas también pueden aprender el procesamiento del lenguaje natural, máxime si

tenemos en cuenta que se pueden automatizar, con un relativo margen de adecuación o calidad, aquellos procesos o tareas en que se den dos condiciones: 1) que las tareas se puedan describir por una secuencia perfectamente definida de acciones elementales y 2) cuando esas tareas se deban repetir muchas veces; ambas condiciones se dan en los procesos de indización, por ello, son perfectamente automatizables. El lenguaje refleja y contiene infinitas posibilidades del pensamiento humano, mientras que las estructuras formales que son los modelos con los que puede operar el ordenador son de naturaleza finita. Una palabra es más que la secuencia de las letras de su significante, a causa del significado que se asocia a estas y de su relación con otras palabras y con el contexto que las rodea. Podríamos explicarlo de una manera un tanto metafórica, que las relaciones que contiene un significante con su significado denotativo y connotativo en cada hablante, son como una *nube* que cuelga de cada elemento del texto y que le parece distinta a cada persona, y el ordenador, no procesa esa *nube*, lo que hace es transformar las cadenas de caracteres.

Con todo lo indicado hasta ahora, podemos decir que nos encontramos en un momento de transición, donde la indización tradicional realizada manualmente para el análisis de contenido de documentos en formato impreso, convive con la indización automatizada destinada al análisis masivo de información textual en formato electrónico.

La indización consiste, pues, en recorrer el documento para comprender y abstraer su magnitud significativa, de tal forma que dé como resultado una representación sintética de su contenido. Esta tarea compleja, exige conocimientos científicos, la comprensión del lenguaje natural y de la lengua del texto y un dominio práctico de un lenguaje documental (sea tesauro, sea lista de encabezamientos o lista de descriptores), además de una capacidad de análisis y síntesis. Todas estas exigencias que podemos estimar para una buena indización pueden concurrir o no en un indizador humano, pero son las que debemos exigirle a un sistema de indización automatizada. Todo análisis semántico de un texto científico es una operación eminentemente intelectual que exige una doble competencia, primero en el plano de la lengua y también en el plano del pensamiento científico, y la máquina debe ser instruida de la misma manera en ambos órdenes de competencia

Los distintos modelos de indización automatizada irán, como veremos a continuación, de una mera extracción en lenguaje natural, donde la palabra se entiende como objeto, pasando por una indización por tratamiento lingüístico sobre un vocabulario abierto, a una indización "inteligente" por conceptos, donde los sistemas de indización y búsqueda se erigen como una verdadera herramienta de búsqueda y recuperación documental.

Modelos de indización automatizada y lenguaje natural

En todos los estudios genéricos –como este– sobre indización automatizada se realizan distintas aproximaciones para caracterizar o tipificar los modelos de indización automatizada, atendiendo a diversos criterios: uno de los más habituales es el criterio evolutivo,[3] en tanto que al ser la indización automatizada un campo de investigación creciente se trata de primar más los avances de esta técnica informatizada de análisis de contenido que la tendencia profética que trate de discernir el futuro de estos sistemas; otro de los criterios más seguidos es el que se fundamenta en método de extracción

terminológica, que distingue fundamentalmente los métodos de extracción lingüísticos de los no lingüísticos, donde los métodos lingüísticos abarcan todas las técnicas derivadas del PLN y los no lingüísticos el resto de las formas de extracción del vocabulario de corte estadístico, probabilístico e incluso, bibliométrico[4] o informétrico; otro de los parámetros que se tienen en cuenta para estudiar los sistemas de indización automatizada es la parte del documento que indizan, distinguiendo esencialmente, los sistemas que indizan las partes principales del documento (título, resumen)[5] de los que se destinan a indizar el texto completo; finalmente, señalamos un criterio fundamental que aparece en múltiples trabajos: el control del vocabulario, que trata de hacer hincapié en la presencia de lenguajes controlados (tesauros o listas de materias) como elemento de control semántico del sistema de indización automatizada frente a una indización exclusivamente *full-text* [11].

Todos estos criterios utilizados para establecer una clasificación de los sistemas de indización automatizada no son excluyentes, más bien responden a un *continuum* de evolución. Lo más habitual es que a tenor de los cambios y de los avances, los modelos no se suplanten, sino que convivan[6] y se aúnen en un fin común, en este caso, conseguir una indización totalmente automatizada. Por ello, trataremos de incluir todos ellos en lo que hemos decidido llamar *generaciones de indización automatizada*, donde parece primar un criterio evolutivo, por razones de claridad expositiva, pero en realidad no queremos revelar sólo la evolución de los sistemas, sino el papel que ha desarrollado en lenguaje natural en cada uno de ellos. Así distinguiremos:

- Una primera generación de la indización automatizada, donde las palabras se entendían como objetos.
- Una segunda generación donde lo que prima es el análisis lingüístico para la desambiguación de conceptos.
- Y, finalmente, una tercera generación a la que hemos denominado indización "inteligente" en tanto que trata de abstraer no sólo conceptos sino modelos conceptuales fundamentados en bases de conocimiento.

Identificación automática de las entradas: la palabra como objeto

Los primeros índices automáticos, construidos por permutación de los elementos que componen las unidades susceptibles de indización (hasta entonces, sólo palabras) fueron los de tipo *kwic-kwoc*. En los años 60, Luhn,[7] conseguía aplicar la capacidad electrónica de los ordenadores a un campo ajeno al de las matemáticas. Pasó así el ordenador a ser considerado capaz de hacer análisis del contenido de los textos. Pero en realidad comenzaba una larga evolución que se desarrollaría entre la capacidad contable inicial y la reflexión cognitiva a la que aspiran las aplicaciones actuales. Desde el comienzo, los ordenadores se utilizaron para procesar textos, en especial para realizar traducciones automáticas [14], lo que está muy cerca de los usos documentales.

Estos primeros intentos se basaron en la identificación de las palabras que aparecían en títulos[8] de artículos científicos. Para hacerlo se utilizaba una base técnica muy sencilla: las palabras se consideraban como objetos exclusivamente y por tanto, desde su

significante. Para llegar a ser una entrada del índice las palabras pasaban primero por el filtro de un antidiccionario, cualquier palabra que constase en este (palabra vacía) y en la unidad que se debía indizar, se eliminaba, y así, las que permanecían se consideraban significativas y pasaban a ser elementos de indización. En la base de cualquier proceso de indización automática se iba a situar desde entonces un algoritmo, cuyo funcionamiento se puede explicar en tres pasos, según muestra en la figura 2.

Fig. 2. Esquema del funcionamiento del algoritmo [15, p. 131][9]

La obtención por este medio de palabras clave daba como resultado innumerables referencias cuando se manipulaba el texto completo, ya que se alcanzaba una indización no selectiva e indiscriminada, incapaz de diferenciar, para el resultado final, las formas flexionadas de una misma palabra por género y número. Y mucho menos aún de reconocer los sinónimos (de tal forma que se podían dar varias entradas para un mismo significado) ni los homónimos (sumando significados distintos al mismo significante). La única posibilidad de orientación hacia el contenido que cada palabra quería representar venía a través de su presentación en contexto. Determinación esta utilizada desde antiguo en la confección de los denominados índices de concordancias [16], cuyo establecimiento se hacía sabiendo que la posible ambigüedad producida cuando las palabras se presentan aisladas quedaba limitada por un contexto que las definía y explicaba.

Los índices permutados tienen una entrada por cada palabra no vacía del documento o fragmento a indizar. Descomponen, por tanto, en elementos simples las expresiones sintagmáticas. La candidatura a ser palabra de indización se originaba exclusivamente en no haber sido eliminada por la lista negativa y en aparecer como caracteres de estructura independiente entre dos espacios del texto en blanco. El texto en ningún caso es tomado como una composición macroestructural, sino como una sucesión de símbolos.

Una consideración que aminora la diferencia entre la utilización del lenguaje natural sin limitaciones y la deseable regulación se establece al observar que muchos de los intentos hechos para indizar mediante ordenadores se han valido de la información

presentada en los registros bibliográficos para facilitar su tratamiento. Partir de títulos y resúmenes ofrece como ventajas tener que procesar un menor volumen, hacerlo sobre la expresión de las ideas sustanciales y encontrar un vocabulario más representativo y, por tanto, más idóneo. Se utiliza así un recurso heurístico de interpretación sumaria del texto completo, aprovechando estrategias que ofrece el propio texto.

Un paso más en la representación automatizada consistió en hacer cálculo de la frecuencia estadística con que aparecían las palabras. Ya no bastaba simplemente con aparecer en la unidad documental que se indizaría para ser considerado candidato, ahora los términos se seleccionaban si su tasa se situaba próxima a una frecuencia de aparición media, quedando fuera las palabras cuyo umbral era muy alto y también aquellas que lo era muy escaso [4]. La utilización del método cuantitativo es la única manera que permite generar algoritmos que haga a las máquinas entender la lengua [17]. Aún así continuaba siendo una indización morfológica, aunque corregida hacia la pertinencia mediante la limitación de aquellas palabras cuya aparición fuera excesivamente abundante o rara dentro de un texto.[10] Sin embargo, el texto seguía siendo considerado una sucesión de símbolos o caracteres, sin prestar atención a la composición macroestructural. Y por ello, al situarnos aún dentro de una indización por palabras, lo implícito, las materias no nombradas, quedaban sin poderse recoger en los índices.

Podemos decir, no obstante, que esta primera generación de modelos para la indización automatizada, basada en criterios meramente estadísticos o probabilísticos, tiene una importancia significativa: por un lado desde el punto de vista de que son los primeros modelos que surgen como alternativa a la tediosa operación documental de la indización aprovechando el desarrollo de la informática y, por otro, porque son métodos que siguen usándose (bien combinados con otros modelos de base más lingüística para la indización, o bien como herramienta para la extracción de palabras en los procesos de elaboración de lenguajes controlados –tesauros) en áreas específicas del conocimiento.

Progresos hacia la desambiguación: la función de las palabras

Ya en los primeros intentos de los años 50 estaba latente un largo proceso para conocer la estructura sintáctica de las oraciones textuales. A principios de los 70 se iniciaban los modelos de análisis lingüístico que se han perpetuado en la mayoría de los sistemas actuales. Esta nueva generación de sistemas de indización automática, debería valerse del procesamiento del lenguaje natural (PLN) cuyos primeros conatos surgían en aquella época, y que en la actualidad ha conseguido unos resultados que sitúan al PLN en posición para liderar una nueva dimensión en las aplicaciones informáticas del futuro: los medios de comunicación del usuario con el ordenador pueden ser más flexibles y el acceso a la información almacenada más eficiente [19, p. 26].

El objetivo era eliminar la ambigüedad de las palabras filtrándolas a través de cuatro procesamientos, análisis o etapas sucesivas –*parsers* lingüísticos– (Fig. 3) de menor a mayor complejidad. Con ellas se busca comprender realmente el significado de los documentos: a) morfológico-léxico; b) sintáctico; c) semántico y d) pragmático.

a) *Procesamiento morfológico-léxico*: En primer lugar, se realiza una segmentación del *corpus* de textos en unidades menores, procediendo a una verticalización de las

oraciones y asignándoles una serie de identificadores que serán utilizados como puntos de referencia en los diferentes análisis posteriores. Se trata no sólo de identificar las palabras, sino también las formas sintagmáticas, las siglas y las locuciones. Los elementos delimitados se contrastan con los dos diccionarios con los que el sistema trabaja (un diccionario que contiene todas las entradas de una lengua; otro con las locuciones e idiotismos), incluso en los sistemas más actuales, las palabras identificadas son sometidas a un proceso de lematización para alcanzar su forma canónica.[11] Debe advertirse que presenta gran dificultad la captación de los conceptos del texto desde el léxico: en primer lugar, porque las asociaciones de palabras se alejan a veces mucho del sentido que tenían sus componentes originales, lo mismo que sucede con los términos polisémicos donde sólo el contexto determina el significado concreto.

Esta etapa tiene como función principal la de obtener el léxico, componente básico de los posteriores análisis sintáctico y semántico; gracias al analizador morfológico, el análisis estadístico de frecuencias se realizará sobre datos formalizados y unívocos semánticamente.

b) *Procesamiento sintáctico*: Utilizando una gramática y/o diccionarios, se analizan las palabras sintácticamente y se describe la estructura de las oraciones. El análisis sintáctico tiene un doble objetivo: por un lado, permite separar las unidades lingüísticas con sentido simples o compuestas y, por otro, permite desambiguar las categorías gramaticales asignadas por el analizador morfológico[12]y al mismo tiempo enriquecer y autogenerar los diccionarios de aplicación.
Los analizadores sintácticos determinan la construcción de las oraciones localizando la función que cumplen las palabras como sujeto, verbo, complemento (y tipos de complementos) [20].

c) *Procesamiento semántico*: Su objetivo es alcanzar el conocimiento temático de los textos, el significado, por tanto, de sus oraciones. Esta etapa se fundamentará, normalmente, bien en un análisis semántico-léxico –estudio de las relaciones paradigmáticas de significado: este análisis permite agrupar y jerarquizar el contenido del texto a través del reconocimiento nuevamente morfológico y del reconocimiento de sinónimos e hiperónimos–, o/y en un análisis semántico-gramatical –estudio de las relaciones sintagmáticas, en el plano de la frase o, y su significado concreto en el contexto del documento– todo ello con la finalidad de reducir y homogeneizar la información léxica del texto que se pretende indizar.

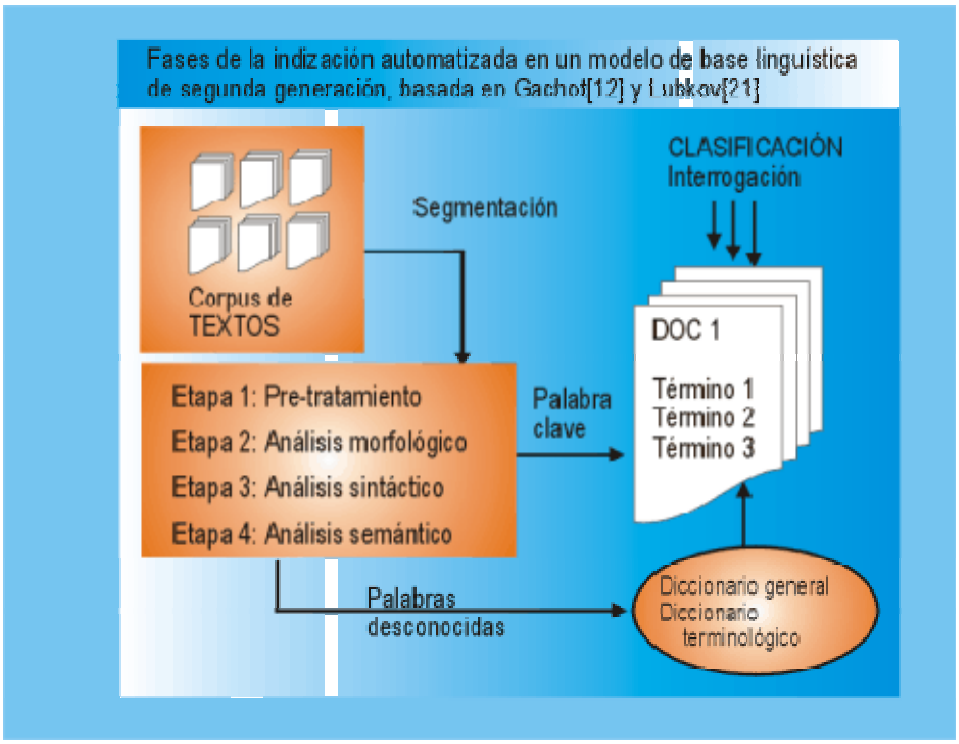

Fig. 3. Fases de la indización automatizada en un modelo de base lingüística de segunda generación, basado en Gachot [12] y Lubkov [21].

Los enlaces dentro de esos esquemas pueden representarse gráficamente mediante estructuras arborescentes que permiten refinar las búsquedas ascendiendo hacia los genéricos descendiendo por los específicos. La base de este análisis se encuentra en los procesos deductivos por los que se establecen inclusiones conjuntivas, llegándose a representar los diferentes dominios conceptuales de un texto.

Para efectuar este nivel del análisis se emplean auténticos tesauros de términos. Los enlaces que estos establecen, ya sea por jerarquías o por asociaciones, permiten precisar o ampliar cada búsqueda dentro de los textos de un campo especializado. No olvidemos que un tesauro contiene los conceptos (y las relaciones que existen entre ellos) mediante los que se representa el conocimiento de un campo científico-técnico. Precisamente la utilización de los mismos tesauros supuso un avance que consistió en que, una vez procesado el texto y extraídos los términos preferentes, pasaron estos a asociarse con dos descriptores de un tesauro. Fue este el inicio de los mapas léxicos donde se representaban los términos del texto y una o varias parejas de términos del tesauro. El ejemplo clásico ha sido el definido por el programa PASSAT *(Programm zür automatischen Selektion vo Stichwörtern aus Texten)* que es el módulo de análisis de textos del software de recuperación de información golem de la empresa informática Siemens.

d) *Procesamiento pragmático*: El análisis pragmático del texto es el más difícil de automatizar ya que implica un conocimiento del mundo real o *semántica de mundo*. Se trata de analizar las relaciones contextuales haciendo uso de algoritmos que permiten comprender el contexto del discurso [22].

Grishman, por ejemplo, advierte en su *Introducción a la lingüística computacional* [23], que una de las mayores dificultades para analizar el contenido de los textos en lenguaje natural es que gran parte de lo significativo está implícito en el discurso. Por eso, algunos de los estudios más avanzados en el desarrollo de *software* para el análisis de contenido, que por ello podríamos incluir en la generación siguiente abocada a una indización *inteligente*, se basan, además de en un análisis puramente semántico, en un Análisis cognitivo discursivo[13] (ACD) y extraen, lo que se denomina *estructura fundamental del significado* (SFS), además de otras técnicas como la constitución de redes semánticas, que veremos en el apartado siguiente.

Hacia una indización inteligente

Las últimas tendencias, que nos permiten hablar de una nueva generación de sistemas de indización automatizada, giran en torno al acceso directo a los documentos a través del procesamiento lingüístico automático y la utilización del lenguaje natural, combinando otras técnicas como el análisis estadístico o la ponderación terminológica. Se busca asegurar la coherencia a la vez que, al utilizar el lenguaje natural, permitir el acceso a los documentos sin formación previa en lenguajes documentales y sin conocer el vocabulario terminológico específico del campo interrogado; esto es, sistemas funcionales que permitan incluir interfaces inteligentes que posibiliten la utilización del lenguaje natural como lenguaje de intercambio de conocimiento entre el documentalista o el usuario final y el sistema. Se trata de integrar todos los modelos y de aprovechar la modularidad en los sistemas para imprimir al ordenador una especie de competencia lingüística y/o cognitiva, teniendo como soporte no sólo bases lingüísticas, sino *bases de conocimiento.*

Podemos decir que en la evolución del procesamiento lingüístico de los documentos ha habido tres momentos marcados por la utilización de otros tantos instrumentos de análisis:

- *Diccionarios* que guiaron el análisis morfológico y el sintáctico utilizando reglas lingüísticas (gramática).
- *Tesauros* que permitieron explicitar las unidades semánticas mediante los enlaces de equivalencia, jerarquía y asociación que existían entre ellos, al aplicar reglas documentales.
- *Bases de conocimiento* que incluso indican los tipos de relaciones que se dan entre los conceptos y desambiguar el contenido del documento.

La gestión del conocimiento, que es la tendencia de todos los sistemas de información actuales, no trata de crear un simple almacenamiento y acceso a la información, sino todo un proceso de manipulación, selección, mejora y preparación de la información, para dotarla de un valor añadido.

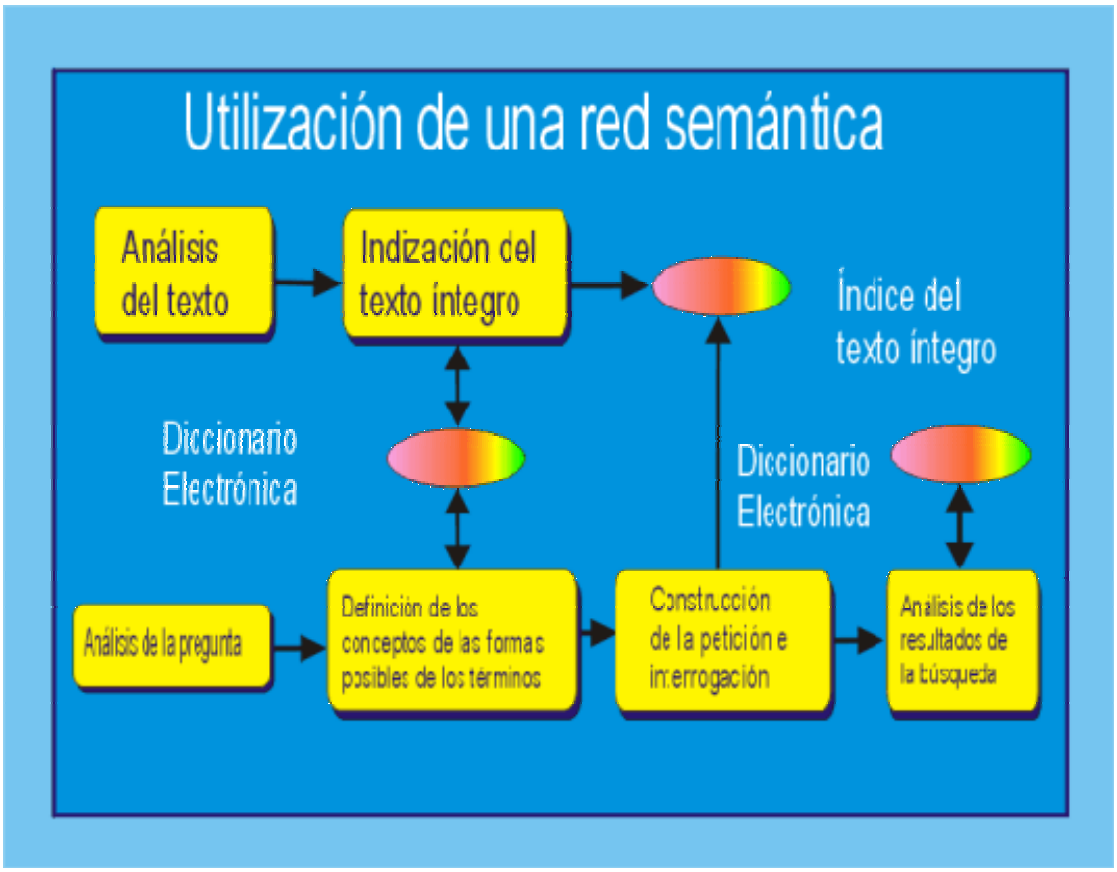

Fig. 4. Utilización de una red semántica

En este sentido la indización automatizada (genéricamente motor de indización y búsqueda) serán un elemento fundamental para la recuperación de información en los nuevos sistemas de gestión del conocimiento y, por ello, se conciben como sistemas de extracción de conceptos, construyendo redes semánticas *input-output* (Fig.4), basadas en bases de conocimiento. Podemos definir un concepto como una representación general y abstracta de un objeto, que permite la recuperación de información por ideas, definidas estas como representaciones distintivas y detalladas de los objetos contenidos en los textos. En estos nuevos motores de indización y búsqueda (*v.gr*. Spirit, y su módulo de análisis semántico Spirit Sense[14] o Tropes[15]) incluidos dentro de software documentales destinados a la GED o a la gestión del conocimiento, podemos atisbar un influjo de las teorías lingüísticas de Saussure y una utilización de la lógica universal aristotélica para construir la semántica del texto y asociar las relaciones del contexto.

Las *bases de conocimiento*, traducción forzada del término inglés *knowledge bases*, según Leloup [25], aparecen, pues, en estos sistemas, como un tesauro enriquecido con información morfológica, sintáctica y semántica, cuyo vocabulario se obtiene del *corpus* de documento de un área del saber. Los textos especializados presentan términos enlazados. Se trata de identificarlos tal como están en los textos, incluso nominalizando los verbos. Como los autores de un campo científico-técnico están al frente de la investigación, su lenguaje está por encima de los controlados y, por tanto, de los que poseen los analistas [26]. Este análisis se fundamenta en el conocimiento que los expertos han depositado en los documentos, es decir, un conocimiento pragmático a

través de la aprehensión de su realidad (*semántica de mundo*). Su aplicación precisa la intervención de la Estadística, la Informática, la Lingüística y la Inteligencia artificial.

En estos sistemas de indización de última generación, se trata, además de asimilar el PLN, de establecer relaciones semánticas desde un hecho con sus causas y consecuencias. Los tesauros ya tenían relaciones de asociación, pero las bases de conocimientos especifican cómo es esa asociación, la representan mediante estructuras arborescentes (generalmente *B-tree*) o en planos. Los términos existen en el texto igual que en los bancos de datos terminológicos, lo que ofrece más posibilidades que el uso de los tesauros que funcionan realmente como diccionarios. El tratamiento lingüístico permite recuperar palabras tanto en su forma canónica como flexionada. Precisamente, al tratar las palabras desde el nivel léxico, su procesamiento se complica, ya que las variaciones terminológicas son innumerables en los textos científicos debido a la inserción de unos términos en otros, a las coordinaciones entre términos, a las variaciones coordinadas y a la morfología derivacional.

Tabla 1. Software de gestión documental destacando el tipo de indización que soporta cada uno de ellos: p-c: palabra clave, *full-text*: texto íntegro, Ln: lenguaje natural

Software	**Empresa**	**Tesauro**	**Tipo de indiz.**	**Módulo web**	**SGBD**	**Cliente**	**Servidor**	**Precio (aproximado)**
Tropes 3.0	Acetic	Sí	p-c, *full text*, Ln	No	Propio	W3.11, 95 y NT	WNT4 WNT5	4 840
Darwin	Cora	Sí	p-c, *full text*, Ln	Sí	--	W3.11, W95-NT, Mac/Os	WNT	--
Cindos 1.1	Chemdata	Sí	p-c, *full text*	Sí	SQLServer Oracle	W95-NT4	WNT4	--
Lexiware 1.5	Erli	Sí	p-c, Ln	No	--	MS IE Netscape	WNT Unix	1 4520
RetrievalWare 6.6	Excalibur	Sí	p-c, *full text*, Ln	Sí	Informix	Browser	WNT Unix	5 2420*
Fulcrum Knowledge Network 2.1	Fulcrum	Sí	p-c, full-*text*, Ln	Sí	Oracle, SQLServer Informix	W3.1, W95-NT	WNT4	--
Spirit 1.6	Technologies Gid	Sí	p-c, full-*text* , Ln	Sí	Propio	Windows Browser	WNT Unix	4 840
Search97 3.10	Verity	Sí	p-c, full-*text*, Ln	Sí	--	--	WNT Unix	--
Zyimage 1.2	Zylab	Sí	p-c, *full text*	Sí	--	W95-NT	WNT4, Novell	4 020

* Precio determinado para 10 accesos.

La última generación de sistemas de indización, busca la representación del contenido utilizando conceptos y algoritmos que dan lugar a nuevas herramientas de software más complejas y dirigidas a la gestión del conocimiento (algunas de ellas se encuentran descritas en la figura 4). Están dirigidas a la indización de textos electrónicos digitalizados; responden a una arquitectura cliente-servidor y a entornos internet/intranet; permiten la indización e interrogación en lenguaje natural; combinan tanto el modelo estadístico (ponderación) con el lingüístico y suelen estar formados por 4 módulos: un módulo de construcción de reglas (canonización), un motor de indización; módulo de cálculo estadístico y un diccionario electrónico o base de conocimiento. Podemos decir con todo, que estos sistemas suponen la asunción del contexto informacional y la solución integrada para indizar el conocimiento electrónico.

Conclusiones

A pesar de que a lo largo de toda la exposición venimos introduciendo algunos puntos de vista sobre el tema, de forma recopilatoria, podemos concluir lo siguiente:

Las investigaciones en torno a la indización automatizada se deben al alto coste de la indización humana (tiempo), al aumento exponencial de la información electrónica, a la proliferación del *full-text*, a la GED, a la informatización de los procesos documentales, a la posibilidad de automatizar los procesos cognitivos y, sobre todo, a la investigación creciente y a los avances PLN. Fruto de estas investigaciones podemos hablar de distintas generaciones de indización automatizada, según el modelo seguido.

La tendencia que siguen las investigaciones en indización automatizada es a integrar todos los modelos y a la modularidad en procesos más simples –análisis estadístico + análisis lingüístico (análisis sintáctico, morfológico y semántico)– de un proceso complejo como es la indización. Aunque son muchos los autores, a los cuales nos adscribimos, que anuncian que el éxito de la indización automatizada vendrá de la mano del desarrollo de las técnicas de procesamiento del lenguaje natural y en el desarrollo de sistemas híbridos y de la inteligencia artificial..., esta modularidad en la que creemos para el desarrollo de la indización automatizada, puede reflejarse también en la necesidad de crear sistemas mixtos que conjuguen el software para el tratamiento del texto completo y la GED, con el software para el PLN.

Las últimas tendencias en indización automatizada han dado lugar a programas específicos para la indización automatizada, pero dentro de software que se destinan a la gestión, almacenamiento y recuperación de información –verdaderos sistemas de gestión electrónica de documentos o sistemas de gestión del conocimiento– donde el módulo de procesamiento/indización (motor de indización) constituye una parte fundamental del sistema (tales programas son, por ejemplo, Search'97, ZyIndex, Excalibur, entre otros). Se tiende, pues a indizar los documentos en formato digital, por medios electrónicos y al acceso directo a los documentos por su contenido a través del

procesamiento lingüístico automático a fin de alcanzar una indización coherente. Al utilizar lenguaje natural, se accedería a los documentos sin formación previa en lenguajes documentales, donde –creemos– el papel del tesauro, como herramienta fundamental para la recuperación de información, no desaparecerá con el desarrollo de las bases de conocimiento, sino que reconvertirá su utilidad más, transparente para el usuario, en los momentos *input-output* del sistema.

El campo de investigación de la indización automatizada y de la recuperación de información es inagotable y se ve magnificado al introducir en él el fenómeno de la gestión de la información en red (internet/intranet). Se trata, pues, de ser receptivos y coherentes con el desarrollo tecnológico de nuestro tiempo, ya que en todo lo que implica extracción de datos *(data mining),* la gestión y la búsqueda del contenido son la próxima etapa, por ello los sistemas de indización "inteligentes" serán el futuro para una verdadera gestión del conocimiento (estructurado o no).

Referencias

1) García Gutiérrez, Antonio. *Estructura lingüística de la documentación, teoría y método*. Murcia, Universidad, Secretariado de Publicaciones, 1990.
2) Bloomfield, L. *Aspectos lingüísticos de la ciencia*. Madrid: Taller de ediciones, 1973.
3) Garfield, E. The relationship between mechanical indexing, structural linguistics and information retrieval. *Journal of Information Science* (18):343-354. 1992.
4) Chaumier, Jaques y Martine Dejean. L'indexation documentaire: de l'analyse conceptuelle humaine à l'analyse automatique morphosyntaxique. *Documentaliste-Sciences de l'Information* 27(6):275-279. 1990.
5) Plaunt, Christian y Barbara A. Norgard. An Association-Based Method for Automatic Indexing with a Controlled Vocabulary. *Journal of the American Society for Information Science* 49(10):888-902. 1998. [También accesibe en: *Papers on Information Retrieval and Autonomous Agents*. Berkeley: University of California, Chris Plaunt's UC Berkeley Web Page, 25 de agosto de 1997. <<http://bliss.berkeley.edu/papers/assoc/assoc.html>>.
6) Charton, Ghislaine. Indexation manuelle et indexation automatique: dépasser les oppositions. *Documentaliste-Sciences de l'Information* 26(4-5):181-187. Juillet-octobre 1989.
7) Gil Leiva, Isidoro y José Vicente Rodríguez Muñoz. De la indización humana a la indización automática. En: *Organización del conocimiento en Sistemas de Información y Documentación.* Zaragoza, Fco. Javier García Marco, ed., 1997, p. 201-215.
8) Jones, K. P.Getting Started in Computerized Indexing. *The Indexer* 15(1):9-13. 1986.
9) Gil Leiva, Isidoro y José Vicente Rodríguez Muñoz. Tendencias en los sistemas de indización automática. Estudio evolutivo. *Revista Española de Documentación Científica* 19(3):273-291.1996.
10) Lisbôa da Silveira Guedes, Vânia. Estudo de um critério para indexãçao automática derivativa de textos científicos e tecnológicos. *Ciencias da Informação* 23

(3):318-326. set-dez 1994.
11) Remize, Michel. Le thesaurus face au texte intégral: une évolution tournée vers l'utilisateur. *Archimag* (112):40-41. Mars 1998.
12)Gachot, Isabelle. Linguistique + statistiques + informatique = indexation automatique. *Archimag* 84:34-37. Mai 1995.
13)Salton, G. The SMART system 1961-1976. Experiments in dynamic document processing. *Encyclopedia of Library and Information Science* 28: 1-28, 1980. Citado por Gil Leiva, Isidoro y José Vicente Rodríguez Muñoz. Tendencias en los sistemas de indización automática. Estudio evolutivo. *Revista Española de Documentación Científica* 19(3):273-291.1996.
14) Locke, William y Donald Booth. *Machine translation of languages*. Cambridge, MIT Press, 1955.
15) Robredo, Jaime. Indexação automatica de textos: uma abordagem otimizada e simple. *Ciencia da Informação* 20(2):130-136. Jul/Dez 1991.
16) Rowley, Jennifer E. *Abstracting and Indexing*. 2nd ed. London, Clive Bingley, 1988.
17) Salton, G., L. Allan and C. Buckley. Automatic structuring and retrieval of large text files. *Communications of the ACM* 37(2):97-108. 1994.
18) Rosenberg, V. A study of statistical measures for predicting terms used to index documents. *Journal of the American Society for Information Science* 22(1):41-50. 1971.
19) Sosa, Eduardo. Procesamiento del lenguaje natural : revisión y estado actual, bases teóricas y aplicaciones. *Information World en Español* 6(12): 26-29. Enero-febrero 1997.
20) Woods, William. Transition network grammars for natural language analysis. *Communications of the AMC* 13(10):591-606. 1970.
21) Lubkov, Michel. L'abc du langage naturel. *Archimag* (103):24-25, abril 1997.
22) Kamp, H. Discourse representation theory: What It is and Where It Ought to go? En: *Natural Language at the computer*, 1988.
23) Grishman, R. *Introducción a la lingüística computacional*. Madrid, Visor, 1991.
24) Ghiglione, Rodolphe, et al. *L'analyse automatique des contenus.* Paris, Dunod, 1998.
25) Leloup, Catherine. *Motores de búsqueda e indexación: entornos cliente servidor, Internet e Intranet.* Barcelona, Ediciones Gestión 2000, 1998.
26) Polanco, Xavier. *Infométrie et ingénierie de la connaisance*. Nancy, INIST-CNRS, 1995.

Bibliografía

Coulon, Daniel, Daniel Kayser. Informatique et langage naturel: présentation générale des méthodes d'interpretation des textes écrits. *Technique et science informatique* 5(2):103-128. 1986.
Gil Leiva, Isidoro. *La automatización de la indización de documentos*. Gijón, Trea, 1999.

Ibekwe, Fidelia. *Traitement linguistique des données textueles pour la recherche des tendances thématiques* [en línea]. Grenoble: Université Stendhal, 1995. <<http://atlas.irit.fr/vsst95/vsst95p8M2.html>>. [Consulta: 11 de mayo de 1999>>.

Indexing Digital Documents it's NOT an Option [en línea]. Texas. University of Texas. 27 de julio de 1997. <<http://fiat.gslis.utexas.edu/~scisco/inel.html>>. (Consulta: el 11 de mayo de 1999)

Moreiro González, José Antonio. Implicaciones documentales en el procesamiento del lenguaje natural. *Ciencias de la Información* 24(1):48-54. Marzo 1993.

Salton, G. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Boston, Addison-Wesley, 1989.

Slype, Georges van. *Los lenguajes documentales de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid, Fundación Germán Sánchez Ruipérez, 1991

Verdejo Maillo, M. F. Comprensión del lenguaje natural: avances, aplicaciones y tendencias. *Procesamiento del lenguaje natural* :5-29. 1994

Recibido: 16 de mayo de 1999.
Aprobado: 15 de junio de 1999.

Eva Méndez Rodríguez
Departamento de Biblioteconomía y Documentación
Universidad Carlos III de Madrid
C/Madrid 126
28903 Getafe (Madrid)
España
Correo electrónico: <<emendez@bib.uc3m.es>>.
<<http://rayuela.uc3m.es/~mendez/home.htm>>.

Notas

[1]Este argumento aparece desde las primeras investigaciones sobre indización automatizada llevadas a cabo en los años 50 (*Cfr.* E. Garfield. The relationship between mechanical indexing, structural linguistics and information retrieval [3]) hasta las más recientes investigaciones de la década de los 90 llevadas a cabo en el INIST (*Cfr.* J. Chaumier et M. Dejean. L'indexation documentaire: de l'analyse conceptuelle humaine à l'analyse automatique morphosyntaxique [4]).

[2]Tal es el caso del trabajo de Plaunt y Norgard, que describen la evaluación de dos algoritmos basados en la técnica de disposición léxica aplicados a 4 626 documentos de la base de datos INSPEC, para crear un *diccionario* de asociaciones entre los ítems léxicos que contienen los títulos, autores y resúmenes y los términos controlados asignados a esos documentos por indizadores humanos, que servirá, en un primer estadio de aplicación del algoritmo, para comparar los encabezamientos de materia asignados de forma automática con los asignados por un catalogador [5].

[3]Este es el enfoque del estudio, por ejemplo, de Isidoro Gil Leiva y José Vicente Rodríguez Muñoz en Tendencias en los sistemas de indización automática [9].

[4]Vânia Lisbôa da Silveira Guedes, es una representante de la corriente brasileña (Río de Janeiro) de aplicación de criterios estadísticos y de leyes bibliométricas –concretamente las *leyes de Zipf* y la *Ley del punto T* de Goffman– a la indización automatizada. *Vid.* Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos [10]. Concretamente, es este artículo, realiza una aplicación de la bibliometría para la indización de un conjunto de textos sobre la mecánica de suelos.

[5]Según Garfield, en facetas del conocimiento muy especializadas (como la Química), un 60% de los términos pertinentes para la indización, están de forma explícita en el título, un 30% está implicado en alguna palabra del título, y sólo el 10% restante se extraía propiamente del texto del artículo [3, p. 344].

[6]Este es el enfoque del estudio, por ejemplo, de Isidoro Gil Leiva y José Vicente Rodríguez Muñoz en Tendencias en los sistemas de indización automática [9].

[7]Vânia Lisbôa da Silveira Guedes, es una representante de la corriente brasileña (Río de Janeiro) de aplicación de criterios estadísticos y de leyes bibliométricas –concretamente las *leyes de Zipf* y la *Ley del punto T* de Goffman– a la indización automatizada. *Vid.* Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos [10]. Concretamente, es este artículo, realiza una aplicación de la bibliometría para la indización de un conjunto de textos sobre la mecánica de suelos.

[8]Según Garfield, en facetas del conocimiento muy especializadas (como la Química), un 60% de los términos pertinentes para la indización, están de forma explícita en el título, un 30% está implicado en alguna palabra del título, y sólo el 10% restante se extraía propiamente del texto del artículo [3, p. 344] disciplinas muy especializadas [3].

[9]1991. La figura muestra el algoritmo de trabajo de los sistemas que extraían el lenguaje natural fundamentándolo en un antidiccionario, según muestra fig. 1, el algoritmo se desarrolla en tres pasos: 1) Las palabras del texto son comparadas con las del antidiccionario; 2) se desprecian aquellas que aparezcan a la par en el texto y en la lista y 3) las que permanecen son consideradas palabras-clave [15].

[10]La utilización de la frecuencia estadística de aparición de las palabras en la representación automática fue ampliamente tratada por Rosenberg [18].

[11]Por forma canónica entendemos la transformación de las formas conjugadas y flexivas en entradas de un diccionario.v

[12]Por esta proximidad en el análisis, algunos modelos de indización de segunda generación prefieren hablar de analizadores morfosintácticos, tratando de realizar un analizador con una gramática particular gobernada por la naturaleza de los textos que se indizan, y cuyo cometido será constituir una serie de modelos que constituyan un repertorio con todas las formas posibles para, a través del análisis flexional y de la lematización, reducirlos a su forma canónica. Esto demuestra que la serie de principios lingüísticos que operan en este tipo de modelos, es constante, pero su orden o

fundamentación teórica es aleatoria.

[13]Sobre este aspecto puede verse Rodolphe Ghiglione *et al. L'analyse automatique des contenus* [24]. Donde se describen las técnicas lingüísticas e informáticas del *software* francés para el procesamiento del contenido textual y la recuperación de información: Tropes de Acetic. Información relativa a este programa, se puede recabar también en la web en <<http://www.acetic.fr/prsentat.htm>>.

[14]Sobre este programa de la empresa T-Gid, puede consultarse a Catherine Leloup. *Motores de búsqueda e indexación: entornos cliente servidor, internet e intranet* [25].

[15]Sobre el funcionamiento y arquitectura del *software* Tropes, resulta muy interesante el libro de Rodolphe Ghiglione *et al* [24].