Artículos

Estudio de algunos aspectos contextuales del discurso médico mediante un módulo automático para el tratamiento documental[1]

Jorge Morato Lara
José Antonio Moreiro González
Manuel Velasco de Diego
Juan Llorens Morillo
José Miguel Fuentes Torres

RESUMEN

Durante los últimos años, la mejora de las herramientas informáticas ha supuesto un notable incremento en la eficiencia de las herramientas documentales. El resultado ha sido un considerable aumento de los documentos electrónicos a texto completo, lo cual conlleva una mayor necesidad de disminuir el ruido en la recuperación documental. Para solucionar el problema, una de las soluciones que está experimentando mayor auge, tanto desde una perspectiva documental como lingüística, es la exploración del contexto mediante el análisis del discurso. En el presente estudio se ha desarrollado una herramienta que con una aproximación multidimensional permite caracterizar determinados discursos dependiendo de las variaciones diafásicas y diamésicas del lenguaje. Los aspectos analizados están centrados en aspectos estilísticos, tipológicos y temáticos del discurso escrito. La herramienta documental desarrollada engloba algoritmos de filtrado y clasificación automática. Así mismo, el vocabulario del MeSH ha sido implementado como herramienta de comparación. El análisis ha sido completado mediante un análisis estadístico multivariante. Se han obtenido diferencias significativas entre los distintos aspectos estudiados, lo cual aconseja el uso de estas aproximaciones para la mejora de las herramientas de documentación automatizada.

ABSTRACT

In this study, a contextual exploration approach is used to analyze texts. There are two related goals of the project; the first one is to study the behaviour of different context parameters when they are faced to automatic filtering and classification tools. The other goal is to improve information algorithms in order to a-priori retrieve contextual information by a set of significant linguistic parameters. The module has been tested using a collection of full-text documents from Medline and Academic Research. The methodology comprises different contexts and type documents. The evaluation of the methodology was accomplished by comparing with the MeSH vocabulary. The text analysis algorithms that have been implemented were the n-grams filter, a hierarchic agglomerative clustering, and finally a co-wording algorithm. Multivariate statistics have showed significant differences among genres and registers. Finally, the discriminate analysis presents promising results to perform contextual analysis automatically.

1 *Los autores quieren hacer patente que el presente estudio ha sido financiado por la Consejería de Educación de la Comunidad Autónoma de Madrid, dentro del proyecto titulado "Aplicación de técnicas informétricas a la construcción automática de tesauros".*

Introducción

Según van Dijk, "una expresión no debería caracterizarse solamente por su estructura interna y significado, sino también en función del acto realizado al producir tal expresión" [1, p. 8]. Es con esta finalidad con la que se va a afrontar en el presente estudio el análisis de ciertos aspectos del discurso que son interesantes desde el punto de vista documental. Por discurso [2] entenderemos la disciplina que estudia el texto y el habla desde todas las perspectivas posibles, es decir, aquella que no intenta sólo clasificar y definir las estructuras, reglas y funciones textuales, si no que también incluye reglas específicas para ciertos contextos y situaciones sociales.

En el discurso existen dos dimensiones a tener en cuenta: la textual, que estudiaría las estructuras del discurso según los diferentes niveles de descripción; y la contextual, que analizaría la dimensión textual de manera conjunta con los diferentes aspectos del contexto, como los procesos cognitivos y los factores socioculturales [3]. Es la faceta estructural la que ocasiona que a principios de la década del 70 surja la lingüística del texto, cuyo cometido propugnaba un análisis del texto que no estuviera limitado a las proposiciones de manera individual [2]. En este sentido, se señalaba que los pronombres, artículos, conjunciones, adverbios y fenómenos, como la presunción o la coherencia, no tienen sentido si no se consideraba el texto en su totalidad. Algunos trabajos sobre todo dentro de la traducción automática y la documentación han surgido a raíz de esta escuela. En concreto entre los estudios que tratan el efecto del número de pronombres en los algoritmos de tratamiento documental estarían los trabajos de Mitkov [4] o de Morato, Llorens, Velasco y Moreiro [5].

Desde que en la década del 50 se publican los primeros trabajos que relacionan lingüística y herramientas documentales [6], no han cesado de aparecer trabajos con nuevas propuestas [7, 8]. No es hasta finales de la década del 80, cuando surgen los primeros estudios encaminados a la realización de un análisis automático del discurso [9]. Simultáneamente, se ha llegado a un razonamiento similar en otras disciplinas con una problemática muy similar a la documental, como en el campo de la traducción. Así, Abaitua, Casillas y Martínez [10] proponen realizar el análisis de textos mediante un estudio previo de la estructura, de la tipología y del registro del texto. Existen cuatro aspectos del discurso que presentan gran interés para el presente estudio: el registro, la retórica, el género y el dominio.

En los siguientes apartados estos aspectos serán analizados con mayor profundidad.

Registro

El concepto clásico de estilo hacía referencia al modo característico en que un concepto es expresado por un determinado grupo de personas, por una persona en particular o en un período concreto. Según este eran los diferentes aspectos del contexto los que determinaban las variaciones en el discurso. Estas variaciones no comportaban un cambio semántico si no que son principalmente de naturaleza léxica y sintáctica. Por ejemplo, podían estar ocasionados por la elección preferente entre términos equivalentes o por una determinada estructura semántica.

Loose [11] ha mostrado en varios trabajos cómo la variación estilística influye en los lenguajes sectoriales. También Karlgren y Cutting [12] han defendido su aplicación, junto con la ecología y la estructura interna del texto, para conseguir una mejora significativa de la recuperación documental.

En el presente trabajo emplearemos el concepto más actual de registro que, aunque es algo más general, está mucho más estructurado y matizado. Halliday [13], con un cierto paralelismo con la idea de estilo, desarrolló el concepto de registro. Este estaba subdivido en tres situaciones según fuera su relación con el entorno: el campo, el tenor y el modo. Aunque todas tienen un marcado interés en el estudio documental, repararemos en el tenor por su relevancia en este estudio. El tenor sería el conjunto de opciones lingüísticas que un determinado sujeto elige según su papel social, de la distancia social con el interlocutor o del momento del discurso en que se encuentre. Como se puede ver, existe un conexión estrecha entre el estilo y el tenor [13] o como es denominado por Lavid [14], por las relaciones interactivas del discurso dentro de cada registro.

Retórica

Según van Dijk [2], la retórica, junto con la estilística y la investigación literaria, sirven para diferenciar los distintos discursos y determinan los efectos específicos de comunicación discursiva. Por retórica se entiende las distintas estructuras opcionales que se suelen emplear en los textos argumentativos para persuadir al receptor del mensaje, por ejemplo, las citas bibliográficas. Las referencias son de hecho la herramienta más utilizada en el análisis bibliométrico de los documentos científicos [15]. Quizás la mayor aportación realizada en la última década sobre este

aspecto, desde la perspectiva de recuperación de la información, sea el trabajo de Lawrence y Bollacker [16]. Estos autores han creado una herramienta que afronta el análisis automático de citas bibliográficas por medio de modelos probabilísticos y reglas heurísticas, sin desvincularlas nunca del contexto en el que fueron formuladas.

Campo de conocimiento

En los últimos años, los experimentos sobre la creación automática de dominios han presentado un gran impulso en disciplinas como la Informática.

La causa principal de este hecho es la necesidad de la reutilización del software [17], lo cual llevaba implícito una mejora en la recuperación de la información contenida en los programas.

Velasco [18] aplicó este método en la generación automática de tesauros, el cual consiste en la aplicación sistemática de una serie de filtros que facilitan los términos con mayor poder discriminatorio. Posteriormente se agrupan los términos en agregados, relacionándolos jerárquica y asociativamente. Polanco, Grivel y Royauté [19] utilizan una metodología similar, pero con el objetivo de crear indicadores bibliométricos por medio de la variación terminológica. Callon, Courtial y Penan [15] también han propuesto su aplicación a la creación de mapas de la ciencia, en los que se presentan una serie de relaciones no jerarquizadas.

Quizás la aproximación más original sea la llevada en los últimos años dentro de la memética [20]. Esta corriente juega con los agregados desde el punto de vista de la ecología de poblaciones, observando los términos como cuasiespecies en el que entran en conflicto factores como el nicho o la competencia. En cualquier caso, en todos estos experimentos se utilizan algoritmos para realizar el análisis de *clusters*. Esta aproximación ha sido frecuentemente utilizada en los estudios de variación lingüística [21]. En estos estudios se puede ver que aunque los experimentos documentales suelen considerar como variables las variantes contextuales, existe un vacío a la hora de su inclusión en los distintos algoritmos. Dado las posibles confusiones con otros conceptos que hacen referencia a entidades más genéricas, como materia o terminología de la disciplina, se ha optado en este trabajo por el término *campo de conocimiento* para referirse al tema concreto del discurso.

Género

Aunque en sus inicios el análisis de género surgió con una finalidad pedagógica [22], por ejemplo, para la enseñanza de lenguas para fines específicos (LFE); hoy día se aplica a una gran variedad de campos. Género y tipología textual son frecuentemente confundidos. Los textos puros no existen en la realidad; por lo que la tipología textual es una abstracción. Así, si, por ejemplo, tomamos el género de artículos científicos, nos encontraremos, frecuentemente, con textos de tipología argumentativa, narrativa y descriptiva. Aún sin olvidar que dentro del registro está incluida la función social que cumple el texto [14], en el presente trabajo se ha considerado en un capítulo aparte. Se ha optado por esta vía siguiendo el criterio impuesto por algunos autores [22], que han aconsejado la utilización de géneros (informes, artículos de investigación, etc.) para distinguir estos escritos, ya muy normalizados, de las elecciones estilísticas más generales que imponen los registros (lenguaje científico, lenguaje burocrático, etc.).

En el análisis del género, destacan los experimentos que durante la última década se han llevado a cabo para la determinación automática de la tipología de varios géneros del discurso. Gilyarevsky, Uzilevsky y Moudrov [23] establecieron un método para diferenciar entre distintos tipos de artículos científico-técnicos de agricultura mediante recuentos de términos en los títulos de cada documento. Con un objetivo similar destaca el trabajo de Haas, Sugarman y Tibbo [24], en este caso se trataría de localizar el vocabulario propio de artículos de tipo "investigación experimental" mediante la generación de dos agregados, de vocabulario empírico y no empírico; los futuros artículos se compararían con estos agregados mediante un coeficiente de similitud. Otros experimentos [12, 25] utilizan una serie más amplia de parámetros junto con un análisis discriminante para diferenciar entre los distintos géneros.

Objetivos

Como se puede observar por lo expuesto, dos aspectos resultan obvios: el primero sería el gran número de factores que avalan el hecho de que las herramientas documentales se comportan de manera distinta ante diferentes contextos, lo que implica que estas herramientas se estarían utilizando de manera imprecisa. Por otra parte, se ha mostrado cómo los estudios adolecen con frecuencia de cierta parcialidad en su visión del problema y se observa la falta de una metodología integradora que permita encarar de una forma eficiente el impacto del discurso en la indización, clasificación y recuperación documental.

En las siguientes páginas, se presenta una aproximación integradora que pretende realizar un estudio pragmalingüístico desde el punto de vista documental. El primero de los aspectos a estudiar sería constatar si para diferentes géneros existe una variación en la clasificación. Se trataría de ampliar el estudio llevado a cabo por Looze y LeMarié [26], en que se compararon los agregados producidos por distintos *corpus* en una misma temática. También, se estudian un conjunto de variables que nos permitan diferenciar entre los diferentes discursos, para lo cual es necesario identificar, previamente, qué variables son las responsables de un mayor porcentaje de varianza en el modelo y cómo estas se interrelacionan entre sí.

Metodología

A continuación se describe el módulo que ha sido desarrollado con el fin de afrontar de una forma integrada el análisis del discurso. Las etapas que se han seguido vienen descritas en la figura 1.

Creación del corpus

Se seleccionó un total de 450 documentos en formato electrónico. La disciplina elegida fue la de documentos biomédicos que versaran sobre pandemias. Se escogieron tres discursos diferentes: documentos científicos, noticias de prensa y artículos de divulgación. Se trataría de estudiar la terminología médica según diferentes registros de texto.

La elección del tema de patologías médicas fue motivada, principalmente, por dos razones, la primera fue la gran normalización presente en el vocabulario, en las superestructuras y en los micromovimientos del discurso médico [27]; en segundo lugar, la gran accesibilidad de los documentos, en todos los registros estudiados, consecuencia natural de la existencia de un gran número de documentos en formato electrónico.

Los documentos procedían de dos bases de datos: MEDLINE y Academic Research.

MEDLINE, con 10 millones de registros, es la más prestigiosa base de datos de medicina a nivel mundial. Esta base de datos está indizada mediante el MeSH [28]. En Medline, se seleccionaron documentos de investigación de las siguientes publicaciones: *New England Journal of Medicine, British Medical Journal, Lancet, Journal of Clinical Investigation* y *Aids Care*. La selección de estas publicaciones estuvo motivada por tener la mayoría de ellas un elevado factor de impacto dentro de sus grupos respectivos, del *Journal Citation Reports* [29]. Hay que hacer notar que *Aids Care* no figura en este catálogo y que la publicación *Journal of Clinical Investigation* se encuentra en un grupo diferente al del resto (está en "Medicine Research Experimental", mientras que el resto está en "Medicine General and Internal").

Academic Search Elite es una base de datos multidisciplinar distribuida por EBSCO. Se trata de una base de datos que contiene un gran porcentaje de revistas evaluadas por expertos. En *Academic Research*, se buscaron artículos de divulgación y periodísticos. Las publicaciones elegidas fueron: *Plant Physiology* (científicos); *US News & World Report, The Economist, Newsweek, Time* (prensa); y *Blood Weekly* (divulgación).

En todas, salvo en *Plant Physiology* y en *Journal of Clinical Investigation*, se realizó una búsqueda por AIDS, hepatitis y scrapie, en 1996. A su vez, se hicieron dos búsquedas paralelas —con el fin de comparar posteriormente los resultados— en el *Journal of Clinical Investigation* sobre bioquímica de proteínas en medicina clínica y en *Plant Physiology* sobre bioquímica de proteínas vegetales. El objetivo era que en la comparación hubiera cierto solapamiento en los temas, aun en campos distintos. Por supuesto, los temas de SIDA y bioquímica de proteínas vegetales serían lo más distantes.

El motivo de elegir estas tres enfermedades era, por una parte, el gran volumen de artículos generados y, por otra, la comparación del componente temporal en los tres campos entre una enfermedad con larga trayectoria en investigación médica (hepatitis), otra

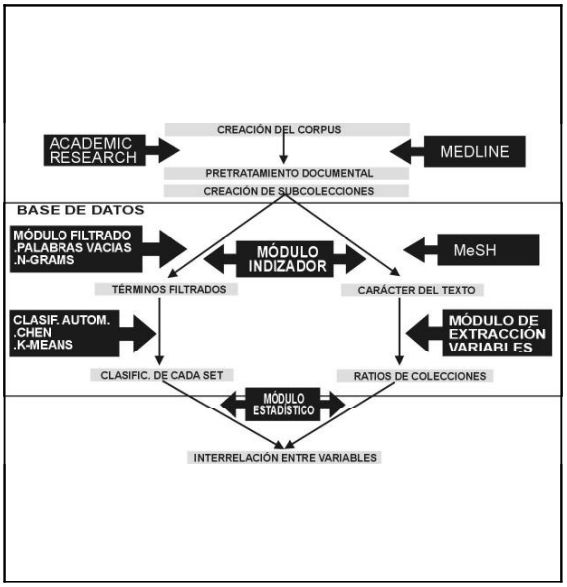

Fig. 1. Metodología de extracción de información.

con una historia más breve (Creufeltz-Jacobson) y, por último, una en pleno auge, el SIDA/HIV.

Pretratamiento documental y creación de subcolecciones

En la medida de lo posible, el ruido fue eliminado mediante la supresión de los caracteres extraños y códigos procedentes de la fuente original. Las características de los documentos, tales como: temática, tipología, autores, instituciones, fuente y palabras clave fueron insertadas manualmente en las propiedades de Microsoft Word del documento.

Por último, con el fin de observar el comportamiento de los diferentes géneros, toda la colección de documentos se fragmentó en varias subcolecciones, según varios criterios:

- Campo de conocimiento (Medicina interna, Investigación médica, Botánica).
- Registro (divulgativo, científico, periodístico).
- Género (actas, notas, editoriales, artículos de investigación…).

Como es lógico todos los documentos estaban dentro de alguna de las categorías de cada criterio. Esta pertinencia fue exclusiva, es decir, un documento no podía ser al mismo un tiempo una editorial y un artículo de investigación.

Base de datos

El MeSH es el vocabulario controlado más extendido en el ámbito médico. Se incorporó el MeSH a la base de datos para tener un referente del comportamiento de las distintas subcolecciones al ser indizados mediante el resultado del filtrado automático por el n-grams. En concreto, se utilizaron dos productos: Medical Subject Headings Tree Structures (MeSH - Tree Structures), con la estructura jerárquica y el Medical Subject Headings -Annotated Alphabetic List: una versión ampliada del MeSH que contiene 18 000 descriptores y 100 000 términos más entre sinónimos y diferentes variantes de cada término, que incluye al mismo tiempo códigos de términos relacionados.

Varios vocabularios específicos se añadieron a la base de datos. Este es el caso del listado de palabras vacías en el que se seleccionó el elaborado para el del programa SMART. En algunos casos, como con los pronombres o los verbos, se añadió la categoría gramatical. También se añadió un listado de términos relacionados con el registro científico-técnico, tomado de Weissberg y Buker [30].

Para el analizador de referencias, se han añadido en la base de datos todos los esquemas que cumplían las referencias y citas en el texto en las publicaciones seleccionadas. El esquema consiste en la especificación de los separadores y el orden en que aparecen los distintos campos de la referencia (autor, título, año y publicación). Un proceso similar se sigue para el marcador en el texto al que hace referencia la cita bibliográfica.

Se identificó la terminología relacionada con el discurso científico. Se realizó mediante dos métodos: por un lado, se recurrió a la propuesta de Leydesdorff [31] consistente en la localización de terminología cuya semántica estuviera relacionada con aspectos tales como la observación, la metodología y la teoría. Por otro lado, términos y frases relacionados con las diferentes secciones del texto se recopilaron de los trabajos de Swales [22], Nwogu [27], Estévez y Martínez-Pelegrín [32] y Skelton [33].

Analizador de citas

Un analizador de referencias fue desarrollado para estudiar alguna de las estructuras retóricas utilizadas. El sistema trabaja de un modo similar al desarrollado en *autonomous citation indexing* [5, 16]. Mediante una serie de datos agregados a la base de datos se localizan, en los párrafos de la bibliografía el autor del documento referenciado, el título, el año y la publicación. A continuación, se localiza en el texto del documento la cita a esos distintos sistemas de citación [22] como, por ejemplo, citas no integrales del tipo: (Sánchez, 91), [SANC 91], [1] o (Sánchez *et al*., 1991), o integrales, como: Sánchez (1990). El sistema localiza en el texto estas citas y las compara con la sección de bibliografía, validando la ocurrencia de la cita en el texto. Estas ocurrencias se emplearán luego en la generación de indicadores.

El analizador tabula todos los autores de las referencias y los vincula con el título de la referencia, año y publicación. Para analizar posteriormente estas publicaciones, fue incluido un listado del *Journal Citation Reports* en la base de datos.

Filtrado mediante n-grams

Se efectuó un filtrado de tipo n-grams con el doble objetivo de aumentar la eficiencia del sistema, al tiempo que se disminuía el ruido [34]. Este algoritmo realiza un filtrado estadístico por medio de la comparación con una serie de cadenas de caracteres. El método es independiente del lenguaje, ya que solo se necesita trabajar con recuentos de las apariciones en el documento y su comparación con un texto más genérico del mismo idioma (*background*). El

background, consistente en la unión de una novela histórica con diferentes artículos de geología, fue elegido tras una larga serie de pruebas. Estos campos de conocimiento fueron seleccionados dado su escaso solapamiento con la disciplina médica. Como ya se ha tratado en trabajos anteriores [18], este método presenta un grave inconveniente derivado de la subjetividad en la selección de un *background* adecuado. El esquema con el que se ha trabajado fue el siguiente: El método se basa en representar el texto mediante secuencias de cadenas de caracteres de un tamaño fijo (n-grams). El número *n*, la longitud de la cadena, suele tomar valores comprendidos entre tres y seis. En este trabajo se ha tomado el valor cinco para poder tener un carácter central en el n-grams. Respecto a la manera en que los documentos eran procesados, los mejores resultados se obtuvieron cuando se analizaban los documentos en lotes de cincuenta documentos solapados cada veinticinco.

Se calculan así las frecuencias de los distintos n-grams observados en el texto. De la comparación de todos los n-grams con el *background* se obtiene un índice que nos señala un conjunto de posibles términos relevantes. A continuación se extraen todas las palabras que contienen los n-grams con mayores puntuaciones, pudiendo formar términos compuestos con palabras próximas.

Análisis léxico

El análisis léxico [35] se realiza con el propósito de transformar las familias de palabras de términos procedentes del n-grams a una misma forma canónica. El módulo de normalización trabaja localizando palabras que no estén identificadas como vacías. A continuación coteja su terminación con una tabla. Cuando la terminación coincide con un registro de la tabla, aquella es sustituida por una terminación normalizada. Se obtiene de esta manera un término candidato. Los términos resultantes se comparan con un vocabulario controlado del mismo discurso que los documentos con los que se está trabajando, validando, o descartando en su caso, el candidato normalizado. Paralelamente, todos los términos de los documentos son normalizados para efectuar una indización mediante el MeSH.

Indización

El resultado del análisis léxico se emplea para indizar los distintos documentos, esto es, para determinar la localización y el número de ocurrencias de cada descriptor en el *corpus* documental. Este dato se guarda en la base de datos para su posterior utilización en la etapa de clasificación y como componente para el cálculo de los indicadores bibliométricos. Paralelamente, como método de comparación del resultado del filtrado y del análisis léxico, se realizó una indización de las subcolecciones mediante el Medical Subject Headings.

Generación de agregados

Se seleccionaron dos algoritmos de clasificación: coocurrencia de términos y k-means. El motivo de elegir esta doble vía está fundado en los trabajos llevados a cabo por Velasco [18] en los que se demuestra cómo estos métodos generan información que, lejos de ser redundante, resulta ser complementaria.

- *Coocurrencia de términos:* Los algoritmos de clasificación por coocurrencias están relacionados con una serie de propuestas realizadas dentro del campo de la Bibliometría para crear mapas de la ciencia [19]. Durante los últimos años, ha existido una tendencia creciente en la realización de estos mapas encaminada a centrarse menos en las citas documentales y más en el cuerpo documental. Estos estudios están englobados bajo la denominación de coocurrencia de términos [15]. La hipótesis en la que está fundamentado el método radica en el supuesto de que si dos términos concurren frecuentemente en varios documentos, la probabilidad de que exista algún tipo de relación semántica entre ellos aumenta. El método de Chen y Lynch [36] trabaja con este algoritmo para generar para cada par de términos una medida del grado de relación.

  Para aplicar este método, se toman los términos generados en la etapa anterior. Posteriormente, se procede a realizar el análisis de coocurrencias para todos los documentos de la colección. Con el propósito de establecer un peso a las relaciones asociativas que existen entre los descriptores tomados dos a dos, se calcula un peso que será comparado con un umbral de significación. Los términos más precoordinados y más específicos alcanzan mayores puntuaciones gracias al modelo de espacio vectorial y la función de semejanza asimétrica empleada.

- *K-means:* Este es uno de los algoritmos de agregados más populares que existen. Aunque existen muchas variantes del algoritmo, una de las más eficientes es el algoritmo convergente de las k-medias de Anderberg. Este algoritmo es parte de la familia de algoritmos de clasificación de centros móviles, donde los centroides son recalculados en cada nueva entrada de datos [37]. Por el procedimiento en que este algoritmo

realiza la agregación es posible la obtención de jerarquías [18].

Partiendo de las ocurrencias documentales de cada término obtenido en el filtrado, este algoritmo comienza por un número *k* de centros temporales procedentes de los conglomerados especificados. A medida que se procesan los casos siguientes se van actualizando iterativamente los centros. Un caso puede sustituir a un centro si la distancia más pequeña del caso al centro es mayor que la distancia entre los dos centros más próximos. De esta manera, se sustituyen sucesivamente los centros que estén más próximos al caso. El resultado final es que todos los casos se agrupan en el conglomerado con el centro más próximo. Velasco [18] ha estudiado las dificultades que supone este método, tanto en la estimación *a priori* del número de agregados, como por la tendencia a crear árboles mal balanceados a causa de la selección inicial de centros.

Ratios bibliométricas y lingüísticas

Se han creado una serie de indicadores que permitieran caracterizar a los distintos géneros [5]. Como mencionan Karlgren y Cutting [12], la elección de los indicadores está motivada por la intuición del investigador y matizada por las posibilidades de programación del sistema. Se seleccionaron dos fuentes principales; por un lado, la adaptación de indicadores procedentes del campo informétrico [38, 39], por otro lado, otros indicadores, que denominaremos lingüísticos, basados en recuentos efectuados sobre características lingüísticas como, por ejemplo, la categoría gramatical.

Las categorías lingüísticas buscan factores tales como la concurrencia en una misma frase de la concurrencia de una cita e identificadores de negaciones o cuasinegaciones [22], asociándose, por último, estas formas al tiempo verbal. Los identificadores negativos, pueden ser de distintos tipos desde adjetivos (*inconclusive, complex, misleading, elusive, scarce*), pronombres o adverbios (*no, little, none, few*), verbos (*fail, lack, overlook*), sustantivos (*failure, limitation*) u otros (*without regard of*).

Es necesario descender a distintos niveles de análisis para valorar estos índices en su justa medida. Tanto las ratio bibliométricas como lingüísticas se calcularon en el ámbito del conjunto de la colección, del documento y de las secciones. Como es obvio, se disminuyó el efecto de la extensión de un determinado texto, calculando el porcentaje respecto al total.

Se calcularon las siguientes medidas: frecuencia y número de palabras y descriptores diferentes; media de palabras por párrafo; legibilidad; tipología documental; número de párrafos; porcentaje de frases negativas con y sin citas; porcentaje de acrónimos; porcentaje de pronombres; y, por último, el porcentaje de verbos en subjuntivo, en pasado, en futuro y en condicional.

En el registro científico, en relación con las referencias en el *corpus* y en cada documento, estarían la frecuencia y el número de referencias; la obsolescencia de las citas y el porcentaje de autocitas.

En cuanto a indicadores procedentes del proceso de indización, se ha calculado el número de descriptores procedentes de información gráfica, procedentes de las leyendas de las tablas y los procedentes de la primera frase de cada párrafo.

Por último, en lo concerniente a los documentos tratados en parejas, se calculó el número de cocitaciones en cada subcolección.

Resultados

Comparación n-grams y MeSH

Para visualizar el comportamiento del módulo de filtrado y compararlo con un producto que nos pudiera ofrecer una visión objetiva, se comparó el resultado con el MeSH. El vocabulario del MeSH tiene una trayectoria de más de treinta años recopilando terminología médica y se ha convertido durante este período en la herramienta con más prestigio mundial. Prácticamente 30% de los 1 800 descriptores filtrados por el n-grams fueron coincidentes con el MeSH. Si bien se observó un mayor número de coincidencias en algunos géneros, en concreto la prensa. Cuando se analizaron los descriptores coincidentes se observó que en el caso de la prensa la coincidencia dependía casi exclusivamente de los términos raíz, es decir, de los más genéricos (Tabla 1).

Si también se incluyen los sinónimos, las diferentes temáticas muestran diferente comportamiento, según el género. Así en el caso del SIDA/HIV, se comprobó que existía una mayor coincidencia en el registro periodístico; sin embargo, al considerar el dominio de la hepatitis, el registro divulgativo consiguió puntuaciones mayores.

Comparación k-means y MeSH

Comúnmente, el algoritmo k-means es utilizado para generar las agrupaciones de objetos que mantienen un

mayor número de características comunes. Estos objetos pueden ser documentos semejantes, usuarios similares, bibliografía coincidente o, como es nuestro caso, terminología relacionada en los documentos de la subcolección.

Tabla 1. Comparación del n-grams y el MeSH

	Porcentaje entre MeSH y n-grams		
	Términos totales n-grams	**Descriptores iguales**	**Descriptores y sinónimos iguales**
Género			
Art. period.	349	21	25
Art. divulg.	830	15	18
Art. invest.	719	12	15
Campo de conocimiento			
Hepatitis	675	18	22
AIDS/HIV	1 585	16	20
Prot. clínicas	549	19	25
Prot. botánicas	56	0	10

Para comprobar el grado de concordancia en el discurso, se realizó una comparación entre los términos agrupados en el MeSH y en el k-means. Estos agrupamientos incorrectos se pueden deber a dos fenómenos: por un lado, a un sobreagrupamiento, es decir, que el número de clases en el k-means sea insuficiente, y se agrupen términos poco relacionados en un determinado nivel de especificidad. Por otro lado, tendríamos un infraagrupamiento, es decir, un número de clases demasiado elevado, en que términos que deberían estar agrupados en determinado nivel de especificidad no lo están.

El comportamiento, conforme aumenta el número de clases en el k-means, comparado con el número de términos incorrectamente relacionados según el MeSH, se puede observar en la figura 2. Como se puede comprobar no todos los géneros muestran la misma dinámica. Se puede observar que algunas campos de conocimiento, como el HIV, y registros, como el divulgativo, presentaron peores trayectorias.

Comparación clasificación por Chen y MeSH

La comparación con el algoritmo de Chen resulta en muchos puntos coincidente con la suministrada en los dos apartados anteriores (Tabla 2). Hay que hacer notar que los agrupamientos que detecta Chen, son principalmente relaciones de tipo horizontal y sinonimias [18], en contraposición al k-means que detecta, también, relaciones verticales o jerarquías. En el MeSH-tree, las relaciones horizontales están menos desarrolladas, lo cual explica los peores resultados en esta comparación. Si bien es reseñable que, en relación con los demás métodos, se registren unos resultados mucho más pobres en lo concerniente a la prensa. En lo referente a los demás apartados hay que destacar que dentro del registro científico, como en los otros experimentos, sí se aprecian mejores resultados cuanto menor es el documento.

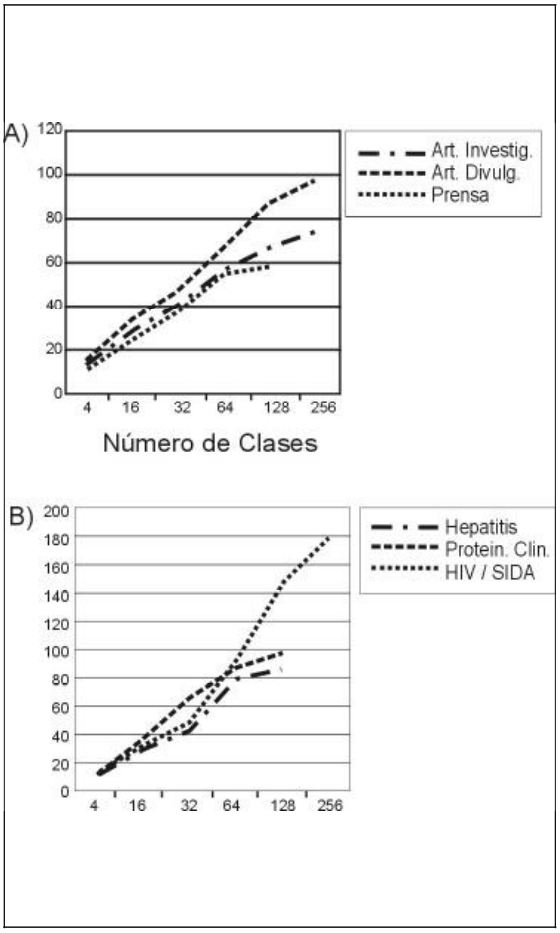

Fig. 2. Comparación k-means y MeSH. Número de términos relacionados en el MeSH y no en k-means, según aumenta el número de clases en k-means. A) Género. B) Campo de conocimiento.

La comparación con el algoritmo de Chen resulta en muchos puntos coincidente con la suministrada en los dos apartados anteriores (Tabla 2). Hay que hacer notar que los agrupamientos que detecta Chen, son principalmente relaciones de tipo horizontal y sinonimias [18], en contraposición al k-means que detecta, también, relaciones verticales o jerarquías. En el MeSH-tree, las relaciones horizontales están menos desarrolladas, lo cual explica los peores resultados en esta comparación. Si bien es reseñable que, en relación con los demás métodos, se registren unos resultados mucho más pobres en lo concerniente a la prensa. En lo referente a los demás apartados hay que destacar que dentro del registro científico, como en los otros experimentos, sí se aprecian mejores resultados cuanto menor es el documento.

Análisis de variables informétricas

Para el análisis de los resultados se realizó, dada la naturaleza de los datos, un análisis multivariante [39], mediante el paquete estadístico SPSS [40]. Si bien fue necesario estudiar los datos previamente con una estadística descriptiva. Se empleó el estadístico de Kruskal-Wallis para analizar si las variables entre los distintos géneros contenían diferencias significativas; se observó que los tres géneros presentan características significativamente diferentes (p,05).

Tabla 2. Coincidencias Chen y MeSH

	Términos en común Chen y MeSH	**Porcentaje respecto MeSH**
Registro		
Divulgación	310	8,9
Científico	306	6,9
Periodístico	80	3,3
Campo de conocimiento		
Proteínas clínicas	100	9,0
Hepatitis	126	7,6
SIDA/HIV	306	6,9
Género		
Actas	66	11,1
Notas	125	9,2
Art. científico	300	7,0
Prensa	66	3,5

Se efectuó un análisis de las componentes principales. El resultado mostró que los componentes que engloban 50% de la varianza están relacionados con el género y el campo de conocimiento. Posteriormente, se observaron únicamente artículos de investigación. En este segundo caso, las dos primeras componentes estaban relacionadas con aspectos tales como las superestructuras y la extensión del texto.

Partiendo de estos resultados, se realizó un análisis discriminante en el que se escogieron las siguientes variables: legibilidad, porcentaje de palabras diferentes, de acrónimos, de pronombres, de párrafos, de terminología científica y de verbos en pasado. El método de agregación de variables fue por pasos, se explican mediante las dos primeras funciones canónicas 98% de la varianza. El resultado se puede observar en la figura 3.

Conclusiones

Se comprobó la existencia de una diferencia estadística significativa entre grupos, lo cual implica que los ratios de clasificación, de asignación de pesos en el filtrado o de indización deben de estar matizados por variables relacionadas con el discurso, el género y el campo de conocimiento. Este hecho dificulta la generalización o mezcla de colecciones en la determinación de estos valores. Por otro lado, los resultados del análisis discriminante muestran

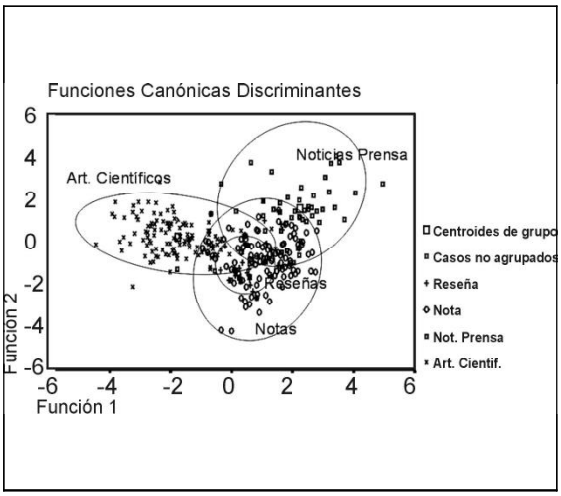

Fig. 3. Análisis discriminante de los distintos géneros.

resultados prometedores en lo referente a la identificación automática del género.

Relacionado con la extensión del documento pueden estar relacionados los resultados obtenidos por los artículos de prensa. Efectivamente, aparte de lo aludido en materia de terminología genérica, los documentos en prensa suelen ser más breves en extensión. En cualquier caso, por el método de filtrado del *n-grams* muchos términos puedan ser los temas colaterales de las noticias de prensa muy raramente alcanzarán el nivel que permita ser identificados como términos valiosos desde el punto de vista discriminatorio.

El diferente estadio de las pandemias seguramente pueda explicar el hecho del diferente comportamiento en sus respectivos géneros. El HIV, con una historia más reciente que la hepatitis, presenta una menor normalización terminológica, lo cual repercute en sus comparaciones con el k-mean y el n-grams.

Como era predecible, los artículos sobre proteínas vegetales no tienen coincidencias con los descriptores del MeSH. Pero cuando se consideraron los sinónimos del MeSH aparecía un diez por ciento de coincidencias. Este hecho seguramente está relacionado con la variación estilística, es decir, los mismos conceptos, por grupos de dominios distintos, son designados de mediante una terminología diferente.

En artículos con la estructura científica característica, se ha comprobado que los indicadores bibliométricos clásicos son realmente responsables de una alta parte de la varianza, y de ahí su poder discriminatorio. Si bien, para su extensión a otro tipo de material es necesario establecer nuevos indicadores.

Desarrollo futuro

Dos aspectos muestran gran interés para el desarrollo futuro: el primero, sería la integración en un módulo de recuperación documental de características relacionadas con variaciones del discurso, principalmente en lo que se refiere al estilo.

También, parece necesario hacer un mayor hincapié en las la variación diacrónica del discurso. Hasta el momento los estudios parecen estar centrados en las características espaciales, pero varios trabajos, principalmente dentro del campo de la Cienciometría parecen señalar que la evolución terminológica de un dominio está en continuo desarrollo [15, 20].

Referencias

1) Dijk, Teun A. van. Texto y contexto: semántica y pragmática del discurso. Madrid, Cátedra, D.L. 1988, 357 p.

2) Dijk, Teun A. van. *La noticia como discurso: comprensión, estructura y producción de la información*. Barcelona, Paidós, 1996, 284 p.

3) Schiffrin, Deborah. *Approaches to discourse*. Oxford, Blackwell Publishers, 1994. 470 p.

4) Mitkov R. The latest in anaphora resolution: going multilingual. *Rev. Proc. Lenguaje Nat.* 23:1-7,1998.

5) Morato, J., J. Llorens, M. Velasco y J. A. Moreiro. Características textuales como medida cualitativa de la información en la generación semiautomática de tesauros. *Rev. Proc. Leng. Nat*. 23:61-68, 1998.

6) Garfield, E. The relationship between mechanical indexing, structural linguistics and information retrieval. *Journal of Information Science* 18(5):343-354. (Presentado en: First Symposium on Machine Methods for Scientific Documentation (Johns Hopkins University, March 1953.)

7) Moreiro González, José A. El resumen científico en el contexto de la teoría de la documentación: texto y descripción sustancial. *Documentación de las Ciencias de la Información* 12:147-170. 1989.

8) Warner, A. The role of linguistic analysis in full-text retrieval. *En* Challenges in indexing electr. text & images. Medford, Learned Inform., 1989. pp. 247-264.

9) Pêcheux, M. Hacia el análisis automático del discurso. Madrid, Gredos. 1978, 374 p..

10) Abaitua, J. K., A. Casillas y R. Martínez. Segmentación de corpus paralelos para memorias de traducción. *Rev. Proc. Leng. Nat*. 21:17-30. 1997.

11) Loose, R. M. Text windows and phrases differing by discipline, location in document, and sysntactic structure. *Inform. Proc. & Manag.* 32(6):747-67, 1996.

12) Karlgren, Jussi y Douglass Cutting. Recognizing Text Genres with simple metrics using discriminant analysis. *Proceedings of COLING 94*, Kyoto,1994.

13) Halliday, M. A. K. Introduction to functional grammar. London, Arnold, 1985, 387 p.

14) Lavid, Julia. Towards a text type taxonomy: a functional framework for text analysis and generation. *Rev. Procesaminento Lenguaje Natural* 16:29-43, 1995.

15) Callon, Michel, Jean-Pierre Courtial y Hervé Penan. Cienciometría: la medición de la actividad científica: de la bibliometría a la vigilancia tecnológica. Gijón, Trea, 1995.

16) Lawrence, Steve, C. Lee Giles y Kurt Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer* 32(6):67-71, 1999.

17) Neighbors, J. Software Construction using Components. Ph. D. Thesis. Department of Information and Computer Science. Univ. California, Irvine, 1981, 176 p.

18) Velasco M. Generación automática de representaciones de dominios. Tesis Doctoral, Universidad Politécnica de Madrid,1998, 241 p.

19) Polanco, X., L. Grivel, L. y J. Royauté. How to do Things with Terms in Infometrics: Terminological Variation and Stabilisation as Science Watch Indicators. *En* Proceedings Fifth Internat. Conf. on Scientometrics and Informetrics. Medford (NJ), Learned Information, 1995. pp. 435-444

20) Best, Michael L. y R. Pocklington. Cultural evolution and units of selection in replicating text. *Journal of Theoretical Biology* 188(1):79-87, 1997.

21) López Morales, Humberto. Sociolingüística. 2ª edición. Madrid, Biblioteca Románica Hispánica, Gredos, 1993, 307 p.

22) Swales, J. M. *Genre analysis: English in academic and research settings.* Cambridge [UK], Cambridge University Press, 1990, 260 p.

23) Gilyarevsky R., G. Uzilevsky y E. Moudrov. An automatic statistical classification of different types of journals. *International Forum on Information and Documentation* 22(3):24-35, 1997.

24) Haas S. W., J. Sugarman y H. Tibbo. A text filter for the automatic identification of empirical articles. *JASIS* 47(2):167-169, 1996.

25) Morato, Jorge. *Análisis de las relaciones cienciométricas y lingüísticas en un entorno automatizado*. Tesis Doctoral. Universidad Carlos III,1999 283 p.

26) Looze, Marie Angèle de y Juliette LeMarié. Corpus relevance through co-word analysis: an application to plant proteins. *Scientometrics* 39(3): 267-280, 1997.

27) Nwogu K. N. The medical research paper: structure and functions. *English Specific Purposes* 16(2):119-138, 1997.

28) Lowe H. J. y G. O. Barnett. Understanding and using the medical subject headings vocabulary to perform literature searches. *JAMA* 13271(14): 1103-8, 1994.

29) SCI. *Journal Citation Reports. A bibliometric analysis of science jounals in the ISI Database*. Editor: Eugene Garfield. Philadelphia, Inst. Scientific Information, 1997, 101 p.

30) Weissberg R. y S. Buker. Writing up research. Englewood Cliffs (NJ), Prentice Hall Regents, 1990, 202 p.

31) Leydesdorff, Loet. Why words and cowords cannot map the development of the sciences? *JASIS* 48(5):418-427, 1997.

32) Estévez N. y P. Martínez-Pelegrín. An approach to the linguistic structures of health science articles. *En* Lenguas para Fines Específicos V: Investigación y enseñanza. Alcalá de Henarés, Servicio de Publicaciones de la U.A.H., D.L. 1996, pp. 301-309.

33) Skelton, J. Analysis of the structure of original research papers: An aid to writing original papers for publication. *Brit. J. General Practice* 44:455-9, 1994.

34) Cohen J. Highlights: Language and Domain-Independent Automatic Indexing Terms for Abstracting. *JASIS* 46(3):162-174, 1995.

35) Frakes W. B. y R. Baeza-Yates. *Information Retrieval. Data Structures and Algorithms*. Prentice Hall PTR. Upper Saddle River, New Jersey, 1992, 504 p.

36) Chen, H. y K. J. Lynch. Automatic Construction of Networks of Concepts Characterizing Document Databases. *IEEE Transactions on Systems, Man and Cybernetics* 22:885-902, 1992.

37) Lelu, C. *Modéles neuronaux pour l'analyse de données documentaires et textuelles*. Ph. D. Université de Paris,1993, 340 p.

38) Osareh, Farideh. Bibliometrics, citation analysis and cocitation analysis. A review of literature. *Libri: International Journal of Libraries and Information Services* 46(3):149-158, 1996.

39) Osareh, Farideh. Bibliometrics, Citation Analysis and Co-Citation Analysis: A Review of Literature II. *Libri: International Journal of Libraries and Information Services* 46(4):217-225, 1996.

39) Flury Bernhard y Hans Riedwyl. *Multivariate Statistics: A practical approach*. London, Chapman and Hall, 1998, 296 p.

40) Voelkl, K. E. y S. Gerber. *Using SPSS for Windows*. New York, Springer,1999, 228 p.

Recibido: 11 de enero del año 2000.

Aprobado: 6 de junio del año 2000.

Jorge Morato Lara

Departamento Biblioteconomía y Documentación
Campus de Colmenarejo. Universidad Carlos III
Av. de la Universidad Carlos III
22 28270 Colmenarejo (Madrid)
España
Correo electrónico: <jorge@ie.inf.uc3m.es>.