GIL LEIVA, I., RODRÍGUEZ MUÑOZ, J.V. Análisis de los descriptores de diferentes áreas de conocimiento indizadas en bases de datos del CSIC. Aplicación a la indización automática. *Revista Española de Documentación Científica*, 1997, vol. 20, nº 2, p. 150-160.

ANÁLISIS DE LOS DESCRIPTORES DE DIFERENTES ÁREAS DEL CONOCIMIENTO INDIZADAS EN BASES DE DATOS DEL CSIC. APLICACIÓN A LA INDIZACIÓN AUTOMÁTICA

Isidoro Gil Leiva* **y José V. Rodríguez Muñoz***

Resumen

Se estudia el valor de los títulos y resúmenes de los artículos científicos como fuentes suministradoras de términos para la indización de los documentos en seis áreas del conocimiento indizadas en las Bases de datos ISOC, IME e ICYT del CSIC. Asimismo, se examina la estructura sintagmática de los términos de indización hallados en el campo "Descriptores", y la posible relación entre el número de descriptores de un documento con la cantidad de páginas del mismo. Para tales fines se seleccionaron las áreas del conocimiento de Biblioteconomía y Documentación, Medicina, Química, Biología, Psicología y Física, y se realizaron seis búsquedas en estas Bases de datos de las que seleccionamos 450 referencias bibliográficas (75 por área) proporcionando un total de 2077 descriptores. El 38,1% de los descriptores asignados a dichos registros aparece en el título, resumen o en el título y resumen a la vez. Como estructuras sintagmáticas descubrimos que el 41,9% de los descriptores son sustantivos, el 32.3% sustantivo+adjetivo, y el 11,8% son sustantivo+de+ sustantivo, quedando solamente un 14% para otras estructuras. Y por último, se han encontrado artículos con escasas páginas y descriptores, documentos amplios y con pocos descriptores asignados, artículos con pocas páginas y una cantidad importante de descriptores, y documentos con un número elevado tanto de páginas como de descriptores.

Se concluye que los títulos cuando no son lo suficientemente precisos, y los resúmenes no están bien elaborados no son fuentes definitivas para la extracción de conceptos; en segundo lugar, que la estructura sintagmática más común es el sustantivo seguido de sustantivo+adjetivo y sustantivo+de+sustantivo; y tercero, que no se aprecia ninguna relación entre el número de páginas de un documento y la cantidad de descriptores asignados.

Palabras clave: Análisis de descriptores, Análisis lingüístico, Análisis estadístico, Indización automática, Bases de datos del CISIC

DESCRIPTORS ANALYSIS ON DIFERENTS KNOWLEDGE AREAS IN CSIC DATA BASE. APPLICATION ON AUTOMATIC INDEXING

Abstract

The value of scientific articles titles and abstracts as sources of terms for documents indexing is studied in relation with six knowledge areas: Library and Information Science, Medicine, Chemistry, Biology and Physics, indexed in the databases ISOC, IME and ICYT of the CSIC. The sintagmatic structures of the indexing terms found in the field 'Descriptors' is also examined, as well the relation between the length of the documents and the number of descriptors it has.

In orden to do this six searches were made in the databases for the six knowledge areas, and 450 bibliographical references were selected (75 for knowledge area), obtaining 2077 descriptors, of these, 38,1% appear in the titles, in the abstracts or in both. With respect to the sintactic structures it was found that 41,9% were 'nouns', 32,3% are 'noun+adjective' groups, and 11,8% are 'noun+noun' groups, with a 14% for other different structures. Lastly, regarding the relationship between length of documents and number of descriptors, all possible combinations were found: short articles with a few descriptors, long articles with a small amount of descriptors, short articles with a important quantity of descriptors, and documents with a high number so much of pages as of descriptors.

The following conclusions can be raised from the data obtained: first, if the abstracts are not well made and the titles are not precise, they are not definitives sources for the extraction of concepts; second, the most common syntactic structures is the 'noun phrase', followed by 'noun+adjective' and 'noun+noun'; third, no significant relation is found between length of documents and number of descriptors assigned to it.

keywords: Descriptors analysis, linguistic analysis, statistical analisys, automatic indexing, CSIC data bases

[*]Departamento de Información y Documentación. Universidad de Murcia

1 Introducción

La indización es una operación compleja pero esta dificultad se torna doble cuando se intenta obtener de forma automática. Mediante la indización automática se pretende que sea un algoritmo el que proponga todos los términos de indización tras el análisis de un documento o algunas de sus partes. Tradicionalmente, en la indización automática se han venido utilizando dos métodos distintos pero a veces convergentes en algunos ensayos, esto es, medios no lingüísticos, iniciados a finales de los cincuenta, y lingüísticos incorporados posteriormente (1).

Cuando se pretende diseñar un sistema de indización automática basado en la extracción de conceptos uno de los planteamientos inmediatos es decidir si las fuentes de las que lograr los términos candidatos a descriptores serán los documentos completos o los títulos y resúmenes de los mismos. Si se elige esta segunda opción cabe preguntarse hasta qué punto los títulos y resúmenes aportan conceptos suficientes para representar el contenido global de ese documento. Otra cuestión a clarificar es si los títulos y resúmenes de documentos científicos -en este caso artículos de revistas- de diferentes áreas del conocimiento proporcionan similar número de conceptos útiles para la indización. Por otro lado, sería interesante averiguar las categorías gramaticales que suelen poseer los descriptores, esto es, si son sustantivos, adjetivos, verbos, preposicones, etc. O si el tamaño de los documentos a indizar está relacionado con la cantidad de descriptores asignados.

En aras de buscar una respuesta a estos interrogantes, se ha abordado el análisis de referencias de artículos de distintas áreas del conocimiento que van desde las humanidades, Biblioteconomía y Documentación o Psicología, pasando por áreas con mayor grado de experimentación como Biología y Medicina hasta llegar a áreas más experimentales como Química y Física, todas ellas indizadas en las Bases de datos ISOC, IME o ICYT del Consejo Superior de Investigaciones Científicas.

2 Material y métodos

De cada una de las áreas seleccionadas se localizaron de una a tres revistas científicas que publicaran trabajos sobre estos ámbitos con la finalidad de que en cada una estuvieran representadas varias subáreas. A continuación, se obtuvieron de las diferentes Bases de datos mencionadas setenta y cinco registros de artículos que cumplieran alguna de estas dos condiciones: que portaran el campo Resumen -imprescindible para los análisis que se deseaban acometer- o bien que se tuviera acceso a la fuente en papel pero que contuvieran el resumen del artículo. Las consultas realizadas se hicieron del siguiente modo:

Búsqueda 1. Para el área de Biblioteconomía y Documentación se hicieron tres consultas a la Base de datos ISOC. En la primera se preguntó por los documentos cuya Fuente de

publicación fuera la *«Revista Española de Documentación Científica»* y se eligieron los veinticinco primeros registros que contenían el campo resumen. Posteriormente, se hizo esta misma operación para localizar documentos cuya Fuente fueran las *«Jornadas Españolas de Documentación Automatiza»* celebradas en 1994 y el *«Boletín de la Asociación Andaluza de Bibliotecarios»* respectivamente.

Búsqueda 2. Para el área de Medicina se dieron estos mismos pasos, pero esta vez en la Base de datos IME se interrogó por documentos publicados por las revistas *«Actas Urológicas Españolas»*, *«Oncología» y «Endocrinología»*. Como en estos registros no venía el campo "Resumen" se tomaron las primeras veinticinco referencias de artículos de cada revista.

Búsqueda 3. La misma operación se realizó para Química en la Base de datos ICYT pero dada la escasez de revistas científicas españolas sobre este área se eligió la única fuente que se tenía accesible en papel que era *«Anales de Química»*.

Búsqueda 4. Para Biología se preguntó a la Base de datos ICYT por estas dos fuentes: *«Anales de Biología»* y *«Monografías de Flora y Vegetación Béticas»*.

Búsqueda 5. En el área de Psicología se utilizó el mismo sistema pero preguntando en la Base de datos ISOC por las revistas *«Investigaciones Psicológicas»*, *«Anales de Psicología»* y *«Anuario de Psicología»*. Estos registros como la mayoría contaba con el campo resumen, se eligieron los primeros veinticinco registros de cada revista que lo contuvieran.

Búsqueda 6. Para Física, debido a la escasez de revistas españolas en este área se eligió *«Anales de Física»* para realizar las consultas en la Base de datos ICYT.

En definitiva, con este proceso obtuvimos setenta y cinco referencias de artículos para cada una de las seis áreas con sus títulos, descriptores y resumen. Y en el caso de que los registros de alguna muestra no contuvieran el campo "Resumen" se tenía perfectamente localizada la fuente en papel para poder acudir a ella. El resumen de los documentos era necesario porque en una de las fases del estudio se debía comprobar qué tanto por ciento de los descriptores asignados a un artículo aparecían desarrollados del mismo modo -ortográficamente hablando- en el título y el resumen de dicho documento. Por tanto, una vez finalizada esta primera fase de búsqueda en las Bases de datos se consiguieron 450 registros con un total de 2077 descriptores. A partir de aquí se comenzó a actuar en cada uno de los objetivos marcados.

El método seguido para descubrir las estructuras sintagmáticas de los descriptores de cada área del conocimiento fue listar todos los descriptores asignados a los setenta y cinco registros, eliminar los repetidos y colocar junto a cada descriptor su categoría gramatical y después, llevar a cabo un recuento de las diferentes estructuras sintagmáticas empleadas. Asimismo, para constatar el número gramatical de los descriptores se utilizaron los listados de los descriptores no repetidos que proporcionó cada muestra. Por otro lado, la impresión de las búsquedas seleccionadas sirvió para establecer lo siguiente: el número total de

descriptores asignados a esos registros, el número medio de descriptores por registro, cuántos eran simples o compuestos, el número mínimo o máximo de descriptores asignados, o la relación o no entre las páginas de un artículo y el número de descriptores.

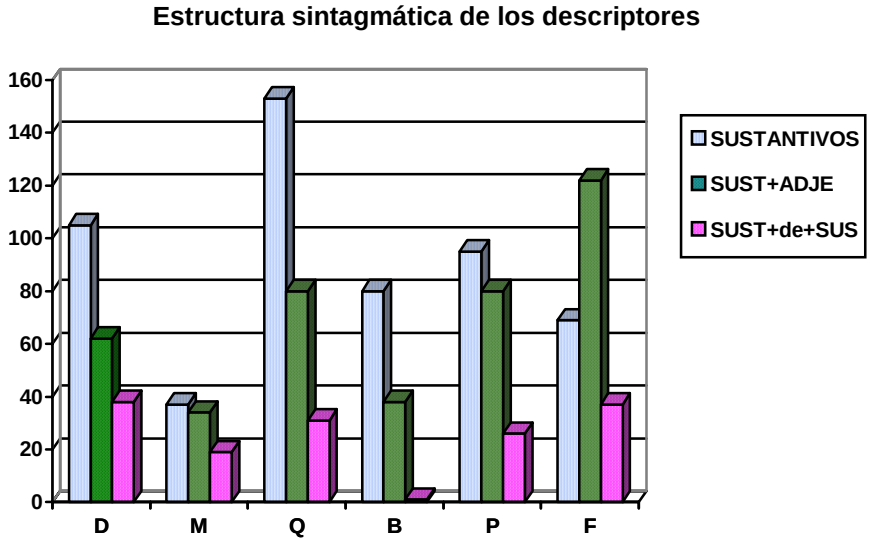
3 Resultados y discusiones

3.1 Estructuras sintagmáticas de las diferentes áreas del conocimiento

Entre las seis áreas del conocimiento analizadas se detectaron un total de 61 estructuras sintagmáticas diferentes. Esta cantidad se reparte del siguiente modo: Biblioteconomía y Documentación (D): 19, Medicina (M): 25, Química (Q): 13, Biología (B): 6, Psicología (P): 21, y Física (F): 20. Lo que supone que la media es 17. Cabe resaltar las apenas 6 estructuras diferentes encontradas en Biología en contraste con las 25 de Medicina.

Tabla I
Estructuras sintagmáticas de los descriptores

Área	Sustantivos	%	Sust+Adje	%	Sust+de+Sust	%	Otras	%
Bib-Doc. (D)	105	45,6	62	26,9	38	16,5	25	10,8
Medicina (M)	37	28	34	25,7	19	14,3	42	31,8
Química (Q)	153	54,6	80	28,5	31	11	16	5,7
Biología (B)	80	51,6	38	24,5	1	1	36	23,2
Psicología (P)	95	40,4	80	34	26	11	34	14,4
Física (F)	69	27,3	122	48,4	37	14,6	24	19,5
Total	539	41,9	416	32,3	152	11,8	177	13,8



Los descriptores cuya categoría gramatical corresponde con los SUSTANTIVOS son los más numerosos, y se manifiestan en todas las muestras estudiadas en primer lugar, excepto en la de Física que aparecen en segunda posición detrás de la estructura SUST+ADJE. La segunda estructura más repetida a lo largo del estudio es SUST+ADJE, que aparece siempre en ese lugar, excepto en la ya mencionada muestra de Física. Y como tercera forma más común SUST+de+SUST que se revela en esta situación en cinco muestras, excepto en la de Biología que este puesto lo ocupan descriptores que denominamos voces latinas. Por tanto, los SUSTANTIVOS acaparan el 41,9 % de los 1284 descriptores no repetidos hallados entre las seis muestras, seguido de la estructura SUST+ADJE (con el 32,3 %) y en tercer lugar, los SUST+de+SUST (con un 11,8 %), lo que supone que el 86 % de los descriptores analizados tiene alguna de estas tres formas, y el 14 % restante lo constituyen un número muy variado de estructuras pero que se manifiestan en muy pocas ocasiones.

Unos ejemplos de descriptores con diferentes estructuras sintagmáticas son los siguientes:

DESCRIPTOR	ESTRUCTURA SINTAGMATICA
INDIZACIÓN	SUSTANTIVO
PUBERTAD PRECOZ	SUST+ADJETIVO
COEFICIENTE DE ACTIVIDAD	SUST+de+SUST
GHRP-6	SIGLA-NÚMERO
ESPECTROSCOPIA RMN	SUST+SIGLA
DICOTYLEDONEAE	VOZ LATINA
BASES DE SCHIFF	SUST+de+NOMBRE PROPIO
TEST DE PATA NEGRA	SUST+de+SUST+ADJE
MÉTODO DE LOS ELEMENTOS DE CONTORNO	SUST+de+los+SUST+de+SUST
CATÁLOGOS DE ACCESO PÚBLICO EN LÍNEA	SUST+de+SUST+ADJE+en+SUST

Para completar lo referido en párrafos precedentes con respecto a las diferentes estructuras sintagmáticas que se han percibido en los registros analizados de las Bases de datos del CSIC se revisaron superficialmente los descriptores de seis tesauros para comprobar algunas de sus estructuras. Aunque en la mayoría de las entradas se repiten como formas más comunes las tres mencionadas anteriormente, se constata en todos ellos una importante variedad. Así en el tesauro Spines (CSIC, 1988) se localizaron más de cuarenta estructuras diferentes como "CRÉDITO PARA LA I+D" ó "INSTITUCIONES QUE OTORGAN SUBVENCIONES". En Eurovoc (C.E, 1987) más de sesenta "FINANCIACIÓN A MUY CORTO PLAZO", "COMITÉ PARLAMENTARIO MIXTO EEE". En el tesauro de Defensa (Ministerio Defensa, 1991) más de treinta estructuras sintagmáticas dispares "AVIONES DE CAZA Y ATAQUE" "CUERPO DE ESPECIALISTAS DEL EJÉRCITO DE TIERRA". En el tesauro del Empleo (Of. Int. Trabajo, 1991) más de treinta: "ESCALONAMIENTO DE LAS HORAS DE TRABAJO" "SERVICIOS SOCIALES PARA LOS TRABAJADORES". En el tesauro de la Unesco

(Unesco, 1982) casi treinta formas "INDUSTRIAS CON FUERTE DENSIDAD DE MANO DE OBRA" "INCITACIÓN AL ODIO Y LA VIOLENCIA". Del tesauro de Medio ambiente (MOPU, 1990) más de treinta "RECOGIDA Y TRANSPORTE DE RESIDUOS" "CONTAMINACIÓN EN LUGARES CERRADOS"

Por otro lado, cabe señalar que en las dos únicas, pero de interés por ser las primeras, propuestas de indización automática para el español [Valle Bracero y Fernández García, 1983] y [Simón Granda y Lema Garzón, 1990] el método utilizado ha sido la búsqueda y extracción de una serie de estructuras sintagmáticas suponiendo que coinciden con posibles términos de indización. Esta opción no la consideramos la más apropiada por las razones que se detallan a continuación. Si se diseña un algoritmo que localice estructuras preestablecidas, es decir, SUSTANTIVOS, SUST+ADJE, etc, supone que previamente hay que efectuar sobre el texto un análisis morfológico y sintáctico para sonsacar las categorías gramaticales de cada una de las palabras y proceder a su desambigüación en el caso de que fuera necesario. Aun admitiendo que se han atribuido correctamente las categorías gramaticales a cada uno de los términos hay que pasar a extraer los posibles términos de indización de un documento partiendo de sus estructuras sintagmáticas lo que implica que hay que contemplarlas todas, y anteriormente se ha comprobado que hay gran cantidad y variedad de ellas. Incluso considerando que estén recogidas todas las posibles formas de los descriptores de un área determinada como por ejemplo Biblioteconomía y Documentación cuando el programa detecte y extraiga los SUSTANTIVOS, los SUST+ADJE, y los SUST+de+SUST, por señalar sólo las más comunes, algunos de los conceptos obtenidos podrían ser los siguientes:

SUSTANTIVOS: "PAÍS" / "CATALOGACIÓN" / "PRÓLOGO"
SUST + ADJE: "TAREA FÁCIL " / "BIBLIOTECAS ESCOLARES" / "INVESTIGACIÓN COMPLEJA"
SUST +de+ SUST: "NÚMERO DE PÁGINA" / "RECUPERACIÓN DE INFORMACIÓN"

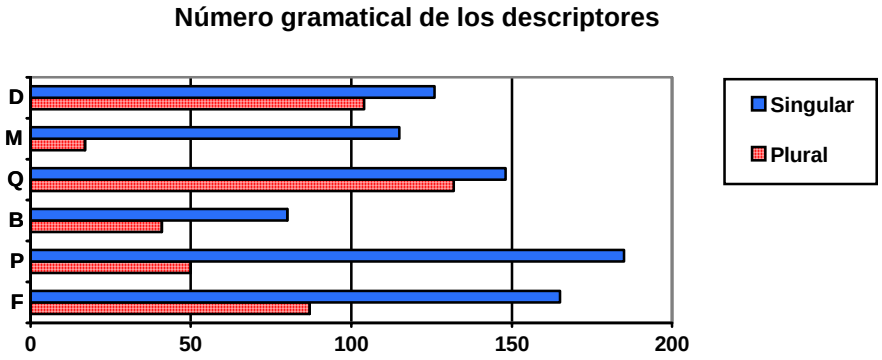
Algunos de los conceptos ("PAÍS", "PRÓLOGO" "TAREA FÁCIL" "INVESTIGACIÓN COMPLEJA", "NÚMERO DE PÁGINA") aun cumpliendo las condiciones sintagmáticas no proporcionan ningún tipo información temática. Pero por el contrario, hay tres conceptos que sí nos transmiten información para inferir que estamos examinando un documento del área de Biblioteconomía y Documentación. Por consiguiente, es necesario el manejo de algún medio para rechazar los primeros cinco términos y validar los otros tres. Este medio posiblemente será una lista de vocabulario autorizado del área a analizar. De este modo, si en última instancia debemos recurrir a una herramienta que autorice o rechace un término o conjunto de ellos, parece, a priori, que no se justifican las complejas operaciones anteriormente descritas de análisis morfológico, sintáctico, localización y extracción de estructuras sintagmáticas. En definitiva, se trata de un proceso complejo tanto de ejecución como de tiempo empleado, por lo que se debe tender a métodos más simples hasta que no se acometan completos tratamientos automáticos de los textos desde el punto de vista semántico y pragmático.

Para finalizar con este apartado se apuntan los resultados del análisis de los descriptores para averiguar el número gramatical de los mismos. Para ello se han tomado de nuevo los 1284 descriptores no repetidos.

Tabla II
Número gramatical de los descriptores

Área	Singular	%	Plural	%
Bibl. y Doc	126	54,7	104	45,2
Medicina	115	87,1	17	12,8
Química	148	52,8	132	47,1
Biología*	80	66,1	41	33,8
Psicología	185	78,7	50	21,2
Física	165	65,4	87	34,5

* Si se suman los datos de esta muestra se comprobarán que faltan 34 descriptores para que sumen los 155 descriptores no repetidos. Estos 34 descriptores son voces latinas por lo que se ha preferido no tenerlos en cuenta, así las operaciones se han realizado sobre 121.



En todas las muestras predominan los descriptores en singular, lo que se traduce en que halla casi el doble (63,7 %) que en plural (32,1 %), lo que parece en principio razonable ya que las normas así lo establecen, esto es, el uso de la forma singular frente a la plural. Asimismo, se observan considerables igualdades o diferencias, ya que en la muestra de Biblioteconomía y Documentación encontramos que los descriptores en singular son 126 mientras que los plurales son 104. Sin embargo, en la muestra de Psicología la diferencia entre descriptores en singular (185) y plural (50) es más que significativa.

3.2 Análisis numérico de los descriptores

a) Descriptores asignados por área de conocimiento

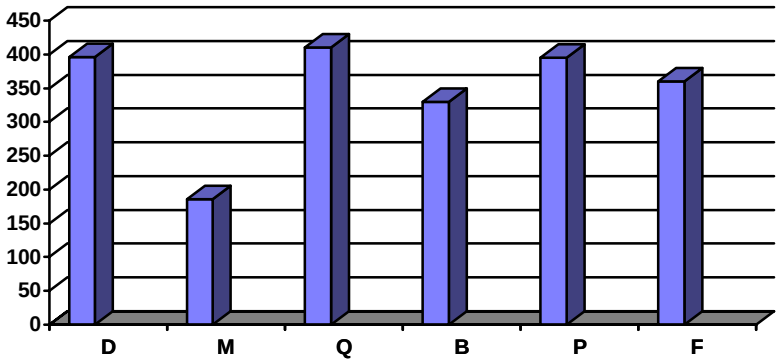
En la Tabla III se observan las diferencias en cuanto al número total de descriptores asignados por área. Si bien se puede advertir cierto paralelismo númerico en casi todas las muestras, en Medicina se aprecia cierta desviación.

Tabla III

Recuento de los descriptores

Área	Nº registros analizados	Nº descriptores asignados	Media descriptores por registro
Bib-Doc.	75	396	5,2
Medicina	75	186	2,4
Química	75	410	5,4
Biología	75	330	4,4
Psicología	75	395	5,2
Física	75	360	4,8

Descriptores por área de conocimiento



b) Media de descriptores por registro

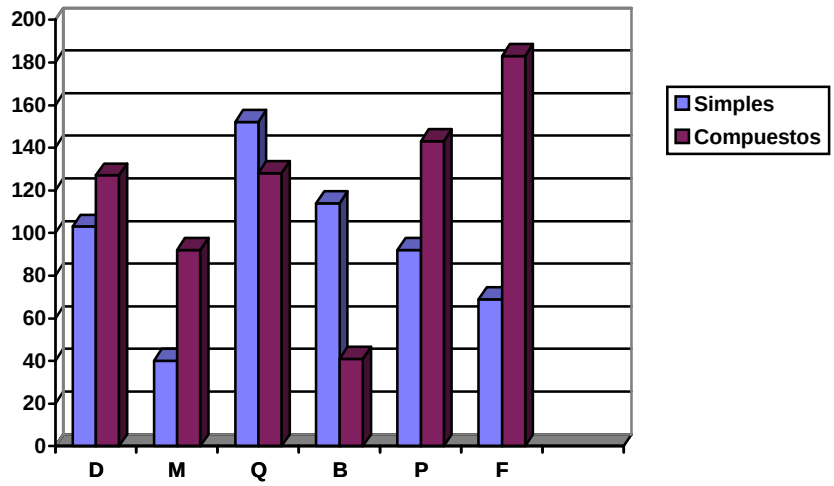
En general, tampoco se produce una gran diferencia entre el número medio de descriptores asignados a cada registro bibliográfico en las Bases de datos consultadas, cuya media en las seis muestras es de 4,5, siendo Química la que tiene la media más alta con 5,4 descriptores y la más baja Medicina con 2,4. Estos datos están en concordancia con los observados en el apartado anterior dado que una mayor asignación de descriptores tiene necesariamente como correspondencia una media más elevada. Palma Villalón [1995, p. 231] recoge los resultados de un estudio de Sievert y Verbeck [1987] en donde comparan la indización de las Bases de datos LISA y ERIC. Estos autores expresaron que el número medio de conceptos utilizados en la indización de los documentos en ERIC es de siete (ocho descriptores), mientras que en LISA la media es de cinco y medio (ocho descriptores).

c) Tipología de los descriptores

En cuanto a si los descriptores utilizados son simples o compuestos hay que mencionar que en dos muestras (Química y Biología) la mayoría de los empleados son simples, mientras que en las cuatro áreas restantes los compuestos superan a los simples. Y en relación a las diferencias más importantes destacan Física con 183 descriptores compuestos frente a 69 simples, y por el contrario, Biología cuenta con 114 simples por 41 compuestos.

Tabla IV
Descriptores simples y compuestos

Área	Nº de descriptores simples	Nº de descriptores compuestos
Bib-Doc.	103	127
Medicina	40	92
Química	152	128
Biología	114	41
Psicología	92	143
Física	69	183
Media	95	119



d) Número mínimo y máximo de descriptores

El número mínimo de descriptores asignados por registro suele ser de dos excepto en Medicina que se han localizado 16 referencias de artículos con un solo descriptor. Por el contrario, el máximo varía desde los doce descriptores hallados en dos registros de Psicología hasta los seis en la muestra de Medicina, en todo caso podemos señalar que no existe una homogeneidad en el número de términos de indización asignados, lo que nos lleva a pensar que no hay criterios generales, en las diferentes fuentes seleccionadas, a la hora de indizar.

Tabla V
Nº mínimo y máximo de descriptores

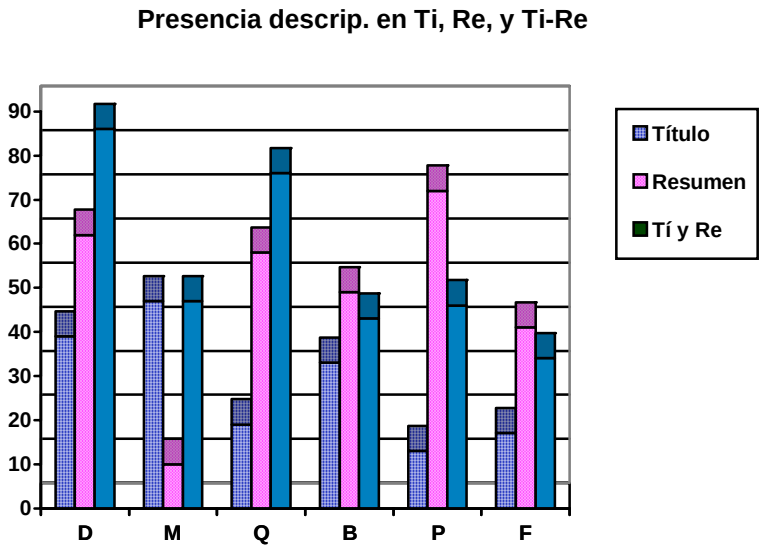
Área	Nº mínimo de descriptores	Nº máximo de descriptores
Bib-Doc.	2	10
Medicina	1	6
Química	2	9
Biología	2	9
Psicología	2	12
Física	2	8

3.3 Presencia de los descriptores en el título y resumen de los documentos

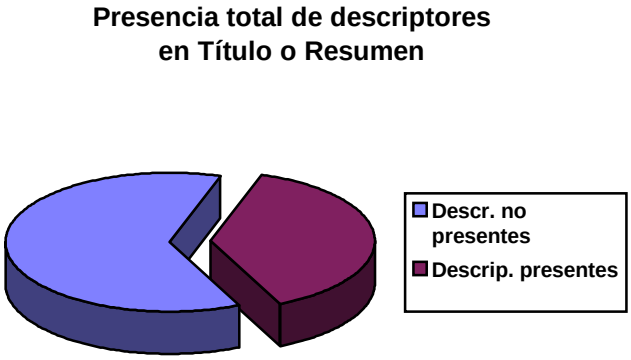
En esta fase del estudio se pretendía hallar qué cantidad de descriptores detectados en el campo "Descriptores" de las Bases de datos analizadas aparecen idénticamente desarrollados bien en el título, en el resumen o en ambas fuentes. Cuando los términos que en dicho campo vienen en plural y en el título o resumen los encontramos en singular se aceptaban como válidos a la hora del recuento, y como es obvio también a la inversa. Tras una tarea laboriosa de comparación de los 2077 descriptores con los correspondientes títulos y resúmenes comprobamos que los descriptores asignados a un documento se exteriorizan más veces en los resúmenes que en los títulos, si bien se ha encontrado una excepción en la muestra de Medicina puesto que de los 186 descriptores asignados, 47 aparecen en el título mientras que tan sólo 10 vienen en el resumen y otros 47 aparecen a la vez en el título y en el resumen.

Tabla VI
Presencia de descriptores en el Título, Resumen y Tí y Re

Áreas	Total de descriptores asignados	Descriptores en Título	Descriptores en Resumen	Descriptores en Título y Resumen	Total descriptores en Título o Resumen	%	Descriptores no presentes en Título ó Resumen
Bib-Doc.	396	39	62	86	187	47,2	209
Medicina	186	47	10	47	104	55,9	82
Química	410	19	58	76	153	37,3	257
Biología	330	33	49	43	125	37,8	205
Psicología	395	13	72	46	131	33,1	264
Física	360	17	41	34	92	25,5	268



Como se puede comprobar en las gráficas precedentes las áreas con una menor presencia de descriptores en el título o resumen son Física, Química, Psicología y Biología. En las dos primeras áreas se esperaban índices más elevados debido a que tradicionalmente, se ha considerado que, en general, los títulos de los documentos científicos en las Ciencias experimentales expresaban más exactamente el contenido de los mismos. En este caso, es al contrario, ya que, si exceptuamos el área de Psicología, en Física sólo encontramos 17 descriptores presentes en los setenta y cinco títulos analizados, y 19 en los de Química.



En definitiva, el 38,1% de los 2077 descriptores asignados a los 450 registros de las seis muestras analizadas están desarrollados del mismo modo (ortográficamente hablando) bien en el título, en el resumen o en título y resumen a la vez. Y por tanto, no están presentes el 61,9% de los descriptores asignados.

3.4 Relación entre el número de páginas y el de descriptores asignados en un artículo

Algunos autores mantienen que existe una relación entre el número de páginas de un documento y los terminos de indización que se le deberían asignar. De este modo, Slype [1991, p. 21] ha expresado que para un artículo de una revista científica o técnica con cinco páginas le corresponderá, en general de 8 a 12 descriptores. Sin embargo, hemos encontrado que en cada una de las muestras estudiadas se reproducen varias de estas cuatro posibilidades: artículos con escasas páginas y descriptores (en la muestra de Medicina: artículos con 2 páginas y 2 descriptores); documentos con bastantes páginas y pocos descriptores asignados (en Biología: artículos con más de 30 páginas y 3 descriptores); artículos con pocas páginas y una cantidad importante de descriptores (en la muestra de Química: registros con 4 páginas y 9 descriptores); y por último, documentos con un número elevado tanto de páginas como de descriptores (en la muestra de Biblioteconomía y Documentación: artículos con más de 30 páginas y 9 descriptores). Lo que nos lleva a considerar que no se constata ningún tipo de relación entre el número de descriptores asignados a un documento y su número de páginas, al menos en las Bases de datos examinadas.

Asimismo, la inexistencia de relación se evidencia igualmente al verificar que en la muestra de Química los artículos suelen tener de dos a siete páginas -sobrepasando muy pocos esta cantidad- y la media de descriptores asignados a esta muestra es la más alta con 5,4, mientras que en la muestra de Biblioteconomía y Documentación el número de páginas por artículo está entre seis y catorce, pero con un número elevado de artículos con más de veinte o treinta páginas, y en cambio, esta muestra tiene una media inferior con 5,2 descriptores asignados por registro.

4 Conclusiones

De las Bases de datos del Consejo Superior de Investigaciones Científicas ISOC, IME e ICYT se han analizado un total de 2077 descriptores asignados a cuatrocientas cincuenta referencias de artículos científicos de estas áreas del conocimiento: Biblioteconomía y Documentación, Medicina, Química, Biología, Psicología y Física. Y tomando como referencia la indización efectuada en estas Bases de datos se desprende lo siguiente:

1. La categoría gramatical que está más presente en los descriptores es el SUSTANTIVO con casi el 42 % del total, pero surgen dos importantes estructuras sintagmáticas por su número como son SUST+ADJE (32,3%) y SUST+de+SUST (11,8%), lo que significa que el 86% de los descriptores analizados tiene alguna de estas tres estructuras sintagmáticas. Y el 14 % restante lo completa una serie de descriptores con una gran variedad de formas, la mayoría de ellas, presente únicamente una o dos veces en cada muestra

2. El número medio de descriptores asignados a las seis áreas del conocimiento analizadas es de 4,5, poseyendo Medicina la media más baja con 2,4, frente a Química con 5,4 descriptores por artículo

3. De los 2077 descriptores asignados a las cuatrocientas cincuenta referencias de artículos, 792 aparecen desarrollados del mismo modo bien en el título, el resumen o en ambas partes, lo que supone el 38,1% de los descriptores asignados, el resto, es decir, el 61,9% no viene ni en el título ni en el resumen. Se observa, no obstante, que los resúmenes son fuentes más interesantes que lo títulos para hallar posibles términos de indización. Sin embargo, si se desea diseñar un sistema de indización automática y se cuenta con herramientas adecuadas y eficaces, es recomendable analizar el documento completo y no ceñirse solamente a la extracción de los posibles términos de indización del título y el resumen

4. El número de descriptores asignados para representar el contenido de un artículo es independiente de la cantidad de páginas con las que cuente el documento.

Referencias

1. GIL LEIVA, I., RODRÍGUEZ MUÑOZ, J.V. Tendencias en los sistemas de indización automática. Estudio evolutivo. Revista Española de Documentación Científica, 19, 3, 1996, p. 273-291

2. PALMA VILLALÓN, Mª. Técnicas y métodos para mejorar la calidad de la indización y su recuperación en bases de datos documentales de Ciencias Sociales y Humanidades. Actas 5es. Jornades Catalanes de Documentació, 1995, p. 223-239

3. SIEVERT, E., VERBECK, A. The indexing of the literature of line searching: a comparison of ERIC and LISA. Online Review, 11, 2, 1987, p. 95-104

4. SLYPE, G. Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales. Madrid: Pirámide, 1991

5. VALLE BRACERO, A., FERNÁNDEZ GARCÍA, J.A. Automatización de la indización y coordinación de descriptores. Revista Española de Documentación Científica, 1983, 6, 1, p. 9- 16

6. SIMÓN GRANDA, J., LEMA GARZÓN, E. Primeras experiencias sobre el análisis de textos en castellano aplicado a la indexación automática de información. Terceras Jornadas Españolas de Documentación Automatizada, 1990, p. 1255-1270