CONTROL VERSUS NO CONTROL DE VOCABULÁRIO EN SISTEMAS DE ARMACENAMIENTO Y RECUPERACIÓN DE INFORMACIÓN

Grupo de Trabajo sobre Lenguajes de Indización
Centro Argentino de Información Científica
y Tecnologica (CAICYT)

RESUMEN

Se reproduce el panel presentado en el Seminario Regional para Docentes sobre Construcción de Tesauros que se realizó en el CAICYT, Buenos Aires, del 19 al 23 de julio de 1982.

Siete especialistas en información efectuaron una confrontación entre el lenguaje controlado y el lenguaje natural. Después de especificar las características de ambos lenguajes, expusieron sus ventajas y desventajas y los criterios para elegir uno u otro. Señalaron la tendencia actual a usar lenguaje natural en los países desarrollados. Mencionaram los sistemas híbridos y analizaron las posibilidades de compatibilizar sistemas que usan lenguajes diferentes.

La concluión es una reflexión sobre las terminologías como nexo entre ambos lenguajes.

1. INTRODUCCION

En ocasión del Seminario Regional para Docentes sobre Construcción de Tesouros, realizado en Buenos Aires del 19 al 23 de julio de 1982 con el auspicio de la Unesco, se presentó un panel sobre el tema control o no control del vocabulario en sistemas de almacenamiento y recuperación de información.

El panel estuvo integrado por las docentes del Seminario, miembros del Grupo de Trabajo sobre Lenguajes de Indización del CAICYT.

La presentación de las diferentes posiciones sobre el tema suscitó un gran interés en los participantes del Seminario y los motivó en forma muy especial.

La idea de organizar un panel se originó en el deseo de comunicar a otros colegas los intercambios de opiniones producidos en el Seminario Regional sobre Construcción de Tesouros de 1978, del cual fue profesor F. W. Lancaster, y durante la preparación del Conjunto Didáctico "Curso sobre Lenguages de Indización: Construcción de Tesouros."

2. DESARROLLO DEL PANEL

Las especialistas integrantes del panel efectuaron una confrontación entre el lenguaje controlado y el lenguaje natural.

Para que la discusión tuviera el rigor adecuado, se inició el tratamiento del tema con la caracterización de ambos lenguajes. A continuación se reproducen los conceptos expresados.

2.1. *Características del lenguaje controlado y del lenguaje natural*

El lenguaje controlado es un lenguaje documentario, elaborado conscientemente para almacenar y recuperar información, a partir de una selección de términos pertenecientes a un subconjunto del lenguaje natural. La lista normalizada de términos se elabora "a priori" de la indización de los documentos. El control de la terminología asegura que las unidades del lenguaje sean unívocas, es decir, que un símbolo o término no represente más de un concepto y viceversa.

El lenguaje libre o natural es el lenguaje utilizado por los autores de los documentos. Es necesario, sin embargo, establecer la diferencia entre lenguaje libre y natural que, aunque sutil, es significativa en el campo de la información. El lenguaje natural es el lenguaje corriente, el lenguaje hablado. El lenguaje libre es el usado dentro de una disciplina de acuerdo con el consenso de la comunidad científica correspondiente. Es en este sentido que existe un cierto control implícito. Al considerar el almacenamiento y recuperación de la información es más apropiado hablar de lenguaje libre y no de natural. Cuando se utiliza el lenguaje libre con este fin no se cuenta con una lista de términos "a priori" sino que se extraen los términos de los mismos documentos.

Una vez caracterizados el lenguaje controlado y el lenguaje libre, se expusieron las ventajas y desventajas que se presentan con la aplicación de cada uno. Para ello, se tuvo en cuenta la relatividad de cada una de las características consideradas y, por lo tanto, se puso énfasis en aquellas ventajas y desventajas más evidentes en cada caso.

2.2. *Ventajas y Desventajas*

2.2.1. *Ventajas del lenguaje controlado*

a) Tratamiento uniforme y consistente de la información mediante el control de sinónimos y casi-sinónimos.

b) Relación semántica explícita.

c) Coherencia entre la indización y la búsqueda.

d) Estructura jerárquica y de asociación que facilita la indización y la búsqueda.

e) Costos menores en la salida porque el tiempo de búsqueda disminuye.

f) Adaptación a sistemas manuales, mecánicos y automatizados.

2.2.2. *Desventajas del lenguaje controlado*

a) Vocabulario limitado y poco específico: predominio de términos genéricos.

b) Entrada costosa porque el especialista que realiza la indización debe dedicar mucho tiempo y esfuerzo.

c) Manejo dificultoso por parte de los especialistas en el campo temático correspondiente.

d) Incompatibilidad de los lenguajes controlados entre sí.

2.2.3. *Ventajas del lenguaje natural*

a) Lenguaje que usan los especialistas en sus comunicaciones escritas.

b) Facilidad para realizar búsquedas por parte de los especialistas ya que no deben ceñirse a la estructura, los criterios y los términos de un lenguaje controlado.

c) Terminología actualizada porque refleja la que se utiliza en los documentos indizados.

d) Posibilidad de una indización y una recuperación más específicas y más exhaustivas.

e) Entrada menos costosa ya que requiere menor esfuerzo intelectual y, en consecuencia, menor tiempo.

f) Posibilidad de compatibilidad terminológica entre distintos sistemas de almacenamiento y recuperación de información dentro de un mismo campo temático.

2.2.4. *Desventajas del lenguaje natural*

a) Inconsistencia en el tratamiento de la información por falta de estructura jerárquica y de control terminológico.

b) Falta de parámetros para el control de la indización.

c) Dificultad para saber si se recupera todo lo que se ha almacenado.

d) Dificultad para la aplicación de recursos de acierto y precisión.

e) Probabilidad de errores en la recuperación por relaciones falsas o incorrectas.

f) Uso exclusivo en sistemas computarizados.

g) Costos mayores en la salida porque el tiempo de búsqueda aumenta.

h) Dificultad para realizar búsquedas por parte de quienes no son especialistas en el tema.

3. CRITERIOS DE SELECCION

Ante la pregunta sobre cómo proceder para elegir el tipo de lenguaje más conveniente para un determinado sistema, las participantes del panel expusieron criterios que deben orientar la decisión, tal como se reproducen a continuación.

3.1. *Ambito o alcance temático del sistema*

Se destacaron las diferencias existentes entre la literatura científico-tecnológica y la literatura sobre ciencias sociales. En el primer caso la terminología es más unívoca, es decir, a un concepto le corresponde un término, y viceversa. En ciencias sociales son más frecuentes los términos similares o iguales que corresponden a conceptos diferentes, así como distintos términos pueden referirse a un mismo concepto. En consecuencia, en aquellas disciplinas en las que hay un mayor control terminológico es más factible el uso del lenguaje natural. En las áreas que presentam muchos problemas de polisemia se aconseja el empleo de vocabularios controlados.

3.2. *Tipos de documentos*

El tratamiento de la información varía según el tipo de documento que posea el sistema. Si predominan los libros, se podrá elegir un vocabulario más general. Si la mayor parte de la colección está constituida por artículos de revistas e informes, se necesitará un vocabulario con mayor nivel de especificidad.

Las actas de reuniones, las patentes y las normas, así como también la información jurídica, se procesan en la mayoría de los casos a través del lenguaje natural.

3.3. *Volumen de la literatura y tasa de crecimiento de la misma*

El volumen y el crecimiento de la literatura pueden determinar las características de la indización, en combinación con otros factores tales como el personal disponible, el presupuesto y los costos.

Si la tasa de crecimiento de los documentos es muy alta, puede recomendarse la utilización del lenguaje natural para facilitar el procesamiento de los documentos (ver ventajas del lenguaje natural).

3.4. *Características de los usuarios*

Para elegir el lenguaje documentaria más adecuado, y especialmente en los casos de búsquedas en línea, hay que tener en cuenta o quienes tendrán acceso directo al sistema. Si fueran los especialistas en los diferentes campos temáticos, convinene usar el lenguaje natural – es el mismo que utilizan esos especialistas – a fin de facilitar las búsquedas. En cambio, si los especialistas en información son los responsables de buscar en el sistema, será más conveniente elegir un vocabulario controlado.

Otros aspectos para considerar son los requerimientos de los usuarios con respecto a la exhaustividad y a la especificidad en la recuperación de la información (ver características del lenguaje natural y del lenguaje controlado).

3.5. *Número de consultas*

La cantidad de consultas puede incidir en la elección del tipo de lenguaje de indización. Si el promedio es bajo se recomienda disminuir los costos en la entrada del sistema mediante la utilización del lenguaje natural y realizar un mayor esfuerzo intelectual en la etapa de búsqueda. El bajo número de consultas justifica emplear más tiempo en la salida del sistema.

Si el número de consultas fuera alto, convendrá efectuar el mayor esfuerzo en la entrada a través de la indización con un vocabulario controlado. Así se evitarán los "ruidos" en la salida y se lograrán

tiempos más breves para las búsquedas, que permitirán satisfacer las numerosas consultas.

3.6. Política de indización

La determinación de la política de indización está condicionada, entre otros factores, por la decisión acerca de los costos en la entrada y en la salida, y por el equilibrio con los beneficios correspondientes.

Las opciones con respecto a la indización están vinculadas a los objetivos y políticas de la institución. Por ejemplo, si se trata de una base de datos comercial, es probable que se abarate la entrada empleando lenguaje natural y se carguen los costos en la salida, ya que esos costos serán cubiertos por los clientes de la base de datos. En cambio, un organismo estatal de nuestros países tratará de absorber ostos y, por lo tanto, procurará que la salida sea menos dificultosa y más precisa mediante el uso de un vocabulario controlado, es decir, realizando el mayor esfuerzo en la indización o sea en la entrada del sistema.

3.7. Implementación física del sistema

La elección del vocabulario está subordinada a la forma en que será utilizado: manual o automatizada.

En los sistemas manuales deben usarse vocabularios controlados. Algunos de estos lenguajes, tales como los tesouros, permiten pasar a sistemas automatizados.

Si se piensa en la implementación automatizada desde un principio, puede optarse por el leguaje natural. Debe entenderse claramente, sin embargo, que no es posible usar el lenguaje natural si se comienza a implementar el sistema en forma manual.

La discusión sobre los diferentes criterios que deben aplicarse para una buena elección del lenguaje de indización, dio lugar a una explicación sobre la tendencia a usa lenguaje natural, que se manifiesta actualmente en los países desarrollados.

4. TENDENCIA ACUTLA EN PAISES DESARROLLADOS

El empleo del lenguaje natural en los sistemas de almacenamiento y recuperación de información se debe, fundamentalmente, al desarrollo de las computadoras y de las temecomunicaciones.

La mayor capacidad y el menor costo de memoria en las computadoras han permitido crear bases de datos que usan lenguaje natural. Además, es posible almacenar textos completos – obviamente utilizando el lenguaje natural – tal como sucede en los sistemas de información jurídica.

Por otra parte, y con la ayuda de las telecomunicaciones, se han difundido los sistemas en línea, cyas terminales pueden ser fácilmente operables por los especialistas en los diferentes campos temáticos cuando se usa el lenguaje natural para la recuperación de la información.

El productor de las bases de datos, como ya se ha dicho, es el más beneficiado cuando se usa lenguaje natural porque ahorra costos en la entrada de información. El usuario es quien debe cubrir los costos de la salida, los que se incrementan al prolongarse el diálogo durante la búsqueda de referencias pertinentes.

En este momento de la explicación se creyó oportuno invitar a participar en el panel a los especialistas del Proyecto para Consultas en Bases de Datos del CAICYT, a fin de que transmitieran sus experiencias.

5. EXPERIENCIAS SOBRE CONSULTAS EN LINEA A BASES DISTANTES

Una especialista en información en el campo de las ciencias sociales describió las características que, generalmente, presentan las bases de datos en dicho campo.

Con respecto a la entrada, en la mayor parte de los casos son versiones automatizadas de bases impresas. Por tal motivo, adoptan el sistema de la versión impresa y lo enriquecen con la posibilidad de hacer consultas en vocabulario libre, es decir, por cualquier palabra que figure en la referencia bibliográfica, en el resumen o en el campo de los descriptores.

Para la salida, casi todas las bases de datos en ciencias sociales cuentan con un tesauro (ejemplo: "LABORDAC" de la OIT), o están clasificadas de alguna forma (ejemplo: EAI que usa CDU), o facilitan listas de grandes categorías (ejemplo: Sociological Abstracts). En todos los casos se puede trabajar en dos niveles, a saber:

1. a partir de las herramientas documentarias mencionadas, se solicita descriptor o categoría elegidos; de esta manera se obtiene mayor precisión con costo razonable;
2. se puede dialogar empleando vocabulario natural y utilizando palabras-clave no sòlo del campo de descriptores sino también del resumen, título, etc., así se recuperan más referencias pero a un costo mayor.

El ejemplo que se menciona a continuación ilustra los dos niveles en que se puede operar.

Se solicita una búsqueda en línea sobre el concepto Democracia Cristiana (DC) y el sistema puede dar dos respuestas:

1. contestar únicamente con las referencias bibliográficas correspondientes a documentos que hayan sido indizados con el descriptor "DC", o sea documentos en los que este tema ocupa un lugar importante;
2. contestar dando todas las referencias bibliográficas que incluyan el término "Democracia Cristiana", ya sea en la cita bibliográfica, en el resumen o en el campo de descriptores, y como consecuencia dar todas las referencias de 1) y otras referidas a documentos no indizados con "DC" pero que tratan el tema en forma periférica.

En la Argentina se prefiere el nivel 1) ya que por el ato costo de las búsquedas conviene reducir el tiempo de diálogo.

6. CARACTERISTICAS DE LAS BASES DE DATOS

Otra especialista en información en el campo de ciencia y tecnología completó la comunicación de las experineicas en el Proyecto con comentarios sobre todas las bases de datos. Consideró que en casi todos los sistemas computarizados es posible utilizar lenguaje libre, ya que fácilmente se crean índices por cada una de las palabras significativas que aparecen en las citas bibliográficas o en los resúmenes.

Informó que, al revisar las 110 bases de datos bibliográficos contenidos en el sistema DIALOG, se determinó que 80 de ellas tienen campo de descriptores con términos extraídos de un vocabulario controlado; tesauros, encabezamientos de materia, listas de términos permitidos, etc. Las restantes no aplican ningún control al vocabulario usado y sólo pueden consultarse utilizando lenguaje libre. De las 80 bases de datos con campo de descriptores controlados, hay sólo 36 que usan un tesauro estructurado con relaciones jerárquicas y de asociación. En 17 de ellas este tesauro puede consultarse "en línea".

En general, se observa que las bases de datos en ciencia y tecnología son más numerosas y también más elaboradas. Incluyen diversos campos: descriptores, identificadores, secciones temáticas, códigos varios, etc. En el campo de los identificadores, los términos son más específicos que los permitidos en un tesauro ya que se extraen directamente del texto del documento original. A través de ellos se pueden efectuar búsquedas en el nivel de especificidad que necesita el usuario.

La exposición terminó con una referencia a la elaboración de estrategias de búsqueda combinando lenguaje natural y lenguaje controlado (lenguaje híbrido), las que otorgan gran flexibilidad al proceso de recuperación de información y amplían las posibilidades de búsqueda.

La mención de este tipo de estrategia introdujo el tema de los sistemas híbridos, que fue brevemente desarrollado por una de las panelistas.

7. SISTEMAS HIBRIDOS

Los sistemas híbridos son sistemas de información que utilizan vocabularios controlados combinados con lenguaje natural. En ciertas condiciones es convenientes implementar esta clase de sistema. Las más frecuentes son las dos siguientes:

1. periferia amplia (macrotesauro y núcleo con alta tasa de crecimiento pero con buen control terminológico (lenguaje natural);
2. compatibilidad con otros sitemas de información a través del uso del mismo vocabulario controlado (se recupera algún campo com lenguaje natural, por ejemplo título).

A continuación, se analizaron las posibilidades de compatibilizar sistemas que usan lenguajes diferentes en um mismo campo temático.

a) *Lenguaje natural*

Dos o más sistemas que utilicen este tipo de lenguaje podrían ser "naturalmente" compatibles.

b) *Lenguaje natural – vocabulario controlado*

Es relativamente fácil integrar un sistema que usa vocabulario controlado a un sistema que utiliza lenguaje natural. La tarea inversa, es decir, el sistema de lenguaje natural integrado al sistema de vocabulario controlado, ofrece más dificultades.

c) *Vocabulario controlado*

La posibilidad de compatibilizar dos vocabularios controlados depende de las siguientes condiciones:

a) grado de superposición temática;
b) grado de control;
c) grado de estructura jerárquica;
d) grado de precoordinación.

Después de la última exposición del panel, la coordinadora invitó a los participantes del Seminario o intervenir en la discusión del tema.

8. INTERVENCION DE LOS PARTICIPANTES

Un especialista en Archivología manifestó que había encontrado en los tesauros la herramienta común para bibliotecarios y archiveros. Destacó la importancia de lograr la compatibilización de los sistemas a fin de favorecer la cooperación ya que sin cooperación no se podrá resolver el problema del control de la información. Para que varias instituciones puedan establecer acciones cooperativas en los procesos de almacenamiento y recuperación de la información, es necesario compatibilizar los sistemas y, especialmente, los lenguajes de indización que se utilizan. El esfuerzo de esta tarea está justificado por el logro de la cooperación.

Otra intervención puso énfasis en la celeridad del desarrollo tecnológico como obstáculo para la compatibilidad de los sistemas de información, tal como sucde en los MEDLARS I, II y III.

Una integrante del panel destacó la importancia de los problemas terminológicos – tanto a nivel internacional como regional – para obtener la compatibilidad de los sistemas. Se referió a las barreras idiomáticas y a la falta de consenso sobre el significado de los términos dentro de una misma disciplina. Como intentos de solución a esos problemas, mencionó a los bancos terminológicos con registro de términos y sus definiciones. Asimismo, destacó el rol que cumplen instituciones internacionales tales como INFOTERM y su red TERMNET.

Algunos participantes intervinieron para expresar su inquietud ante la importancia que las panelistas habían dado al lenguaje natural. Consideraron que en los países latinoamericanos es más apropiado el uso del lenguaje controlado.

Los integrantes del panel respondieron que estaban de acuerdo con esa observación y aclararon que al hablar de lenguaje natural se había querido señalar la prospectiva de los vocabularios de indización.

Otros participantes señalaron el énfasis que las panelistas habían puesto en la metodología de Lancaster y preguntaron por qué no se recurría al pensamiento de otros especialistas, tales como los pertenecientes a la escuela inglesa (Clasification Research Group).

Se explicó que Lancaster era el autor más conocido para el Grupo de CAICYT por el curso que había dictado en 1978 y porque se habían realizado experiencias de construcción de tesauros en base a sus enseñanzas.

Para dejar aclarados ciertos conceptos sobre el lenguaje natural y a manera de primeras conclusiones, el panel expuso las siguientes ideas.

Las posibilidades de acceso a equipos de computación condicionarán el tipo de lenguaje a elegir. Sólo se puede pensar lenguaje natural cuando se usa computadora.

En América Latina se debe continuar con los vocabularios controlados ya que la automatización es costosa por varias circunstancias: alto costo de los equipos con gran capacidad de memoria, falta de personal especializado, carencia de "software" adecuado, dificultad para adaptar sistemas y programas a nuestra realidad.

Cuando existe un equipo de computación disponible, los servicios de información deben tener en cuenta qué tiempo y qué parte de la memoria podrán utilizar. Esta situación está relacionada directamente con la elección del lenguaje. Si es necesario incluir los resúmenes de los documentos para lograr ura recuperación con lenguaje natural, el archivo debe ser más voluminoso que si se usa vocabulario controlado para realizar búsquedas a través de descriptores.

Para mostrar el acercamiento entre el lenguaje natural y el lenguaje controlado y a manera de conclusión final, la coordinadora expuso la siguiente reflexión.

9. CONCLUSION FINAL

El progreso en todos los campos depende de la comunicación de la información. Una comunicación efectiva sólo puede lograrse si los conceptos – elementos del pensamiento – tienen el mismo significado para todos los participantes del proceso de la comunicación.

Los miembros de las diferentes comunidades científicas han tomado conciencia de este problema y, por talmotivo, en todos los campos del conocimiento se están realizando esfuerzos para establecer terminologías. Esto significa llegar a acuerdos sobre conceptos y términos.

Por otra parte, los tesauros deben tomar como base las terminologías aceptadas tanto por los generadores como por los usuarios de la información.

Las terminologías, por lo tanto, acercan los lenguajes naturales a los lenguajes controlados. Constituyen el nexo entre ambos tipos de lenguajes.

El panel estuvo integrado por las siguientes especialistas: Mónica Allmand, Graciela Carballo, Celia Molina, Mercedes Patalano, María Laura Pesce, Ana María Sanllorenti y Ethel Zítara de Ribezzo, quien actuó como coordinadora.

Participaron también las especialistas del Proyecto Consultas en Bases de Datos del CAICYT, Dominique Babini y María Angélica Porta.

CAICYT
Moreno 431
1091 BUENOS AIRES
ARGENTINA

ABSTRACT

This paper reproduces the panel developed at the Regional seminar for Teachers on Thesauri Construction, held at the CAICYT, Buenos Aires, from 19 to 23 July 1982.

Seven information specialists made a comparison between controlled and natural languages. They detailed the characteristics of both types of languages, stated their advantages and disadvantages and indicated the criteria to be applied in order to choose one or the other. They also pointed out that nowadays developed countries tend to use natural languages. They dealt with hybrid systems and analyzed the possibilities of achieving compatibility among systems which use different types of languages.

Finally, it arrived to the conclusion that terminologies serve as links between both languages.