Algunas grietas semánticas en la recuperación de información: una perspectiva deconstructiva para una solución pragmática

Ibarra Contreras, Rafael
ibarraraf@aol.com
ovlop@servidor.unam.mx

Abstract

[Spanish abstract]
Describe la problemática que enfrentan los nativo hablantes del español al consultar bases de datos de información diseñadas para su consulta en idioma inglés y ofrece una solución pragmática por medio de un algoritmo lingüístico en que se describe la ruta crítica y la ruta ideal.

[English abstract]
Describes the general problems faced by Spanish native speakers when retrieving information from databases designed to be used by English native speakers and offers a pragmatic solution by means of a linguistic algorithm which describes a critical route and an ideal one.

[Spanish keywords]
Algoritmo; bases de datos de información; deconstrucción; entropía; lingüística; nativo hablantes del español (NHE); Recuperación de información (RI).

[English keywords] algorithm; deconstruction; entropy; information databases; information retrieval (IR); lingusitics; Spanish native speakers (SNS).

Actualmente, prevalece la problemática semántica subyacente en la recuperación de información (RI) de bases de datos diseñadas en idioma inglés (BDI) empleadas por nativo-hablantes del idioma español (NHE), lo que representa una inversión lamentablemente considerable en tiempo y en dinero empleados por la falta de resultados rentables.

Seguramente todos los que nos hemos ocupado en buscar y recuperar información en fuentes electrónicas, como las BDI, hemos experimentado cierto grado de agrietamiento, ya sea porque encontramos demasiada, poca, o ninguna información para desarrollar nuestra investigación, saciar nuestra curiosidad o satisfacer nuestras necesidades respecto al quehacer académico o comercial.

En algunas ocasiones, los resultados extremos nos han conducido a prestar poca atención sobre las causas o razones ante la demasía, irrelevancia o de la ausencia de datos; aunque también hemos dedicado algunos momentos a reflexionar el por qué de los resultados obtenidos, ¿qué hice mal?, ¿qué más debo hacer?, preguntas que hacen disparar especulaciones igual de extremas como, "seguramente que soy pionero en el tema", "necesito cambiar de tema de mi investigación porque ya hay demasiado trabajo realizado", entre otras.

Esta clase de rutinas produce un efecto atroz, ante la insuficiencia de soluciones razonadas y prácticas por parte de los proveedores de las BDI en sus sistemas de recuperación de información (SRI), o a falta de propuestas de investigación por parte de los profesionales en las ciencias de la información NHE, es común que, tanto usuarios como bibliotecarios, se desarrollen paulatinamente dentro de un estado rústico de "naturaleza profesional" en que ellos se desenvuelven: un ámbito académico autosuficiente y aislado uno del otro, sin importar los defectos o fallas desde los que se trabaja.

Cada uno de los elementos de este escenario es conducente a la costosa aparición de grietas semánticas, por las que se cuela el sentido, y que exige de la propuesta de una aproximación remedial ahora y, preventiva después, como la que se presenta en este trabajo. Tal tarea resultaría arriesgada si se intentara repetir o mejorar los esfuerzos ya hechos. Este no es el caso.

Desde algunos años antes de iniciar la presente propuesta, se han observado diversos acercamientos para dar respuesta a la entropía y redundancia como producto común de las BDI. La gran mayoría de ellos han sido realizados para resolver el problema desde un punto de visita del nativo hablante del idioma inglés (NHI), cuyos enfoques están centrados desde el punto de vista del SRI. Lo anterior se mostrará con literatura revisada en la BDI *Library and Information Science Abstracts*.

Los esfuerzos en investigación de NHE, documentados en el catálogo de tesis TESIUNAM, ofrecen aspectos descriptivos en cuanto a la modificación del ambiente en el que se trabaja, inducción de aproximación a los servicios que se ofrecen en general, análisis de necesidades, derechos de autor, medida de satisfacción, instrucción bibliográfica, conducta en la biblioteca, hábitos de lectura, personal y usuarios,

evaluación de revistas, procesos de clasificación, “desarrollo de búsquedas efectivas”, descripción de bases de datos – en español, normas ISO en diseño de bases de datos, diseño y elaboración de bases de datos, estructuración semántico-pragmática en sistemas de recuperación de información enfocado a usuarios NHE.

Es con base a estos argumentos en que se decide ofrecer y discutir una perspectiva deconstructiva y una alternativa de solución pragmática que comprende ilustrar las “oposiciones” existentes – inteligencia artificial vs sensibilidad humana, en el ámbito de RI en BDI empleadas por NHE, a través de un esquema que ilustra un breve examen de la interacción del lenguaje empleado por usuarios y bibliotecarios. El objetivo es contribuir a disminuir la vasta brecha de malentendidos desde una postura lingüística crítica, deconstructiva, por medio de la ilustración de la ruta crítica típica infructuosa y la ruta ideal fructífera en la RI. Se describe una serie de factores lingüísticos y humanos opuestos y sus impactos en la RI, iniciando por algunas consideraciones del idioma inglés.

El idioma

Hoy en día, como académicos NHE, enfrentamos los impresionantes resultados y confusiones de la tecnología, relacionada con los campos de la bibliotecología y las ciencias de la información, conocidos como BD. También tenemos que manejar la lingua franca en que este tipo de tecnología se exige: el idioma inglés. En este momento, los datos académicos son presentados y susceptibles de buscarse y recuperarse por impresionantes grupos de prestigiosas BD, que incluyen más de dos o tres mil títulos de journals arbitrados, además de memorias de congresos, coloquios, simposio y libros.

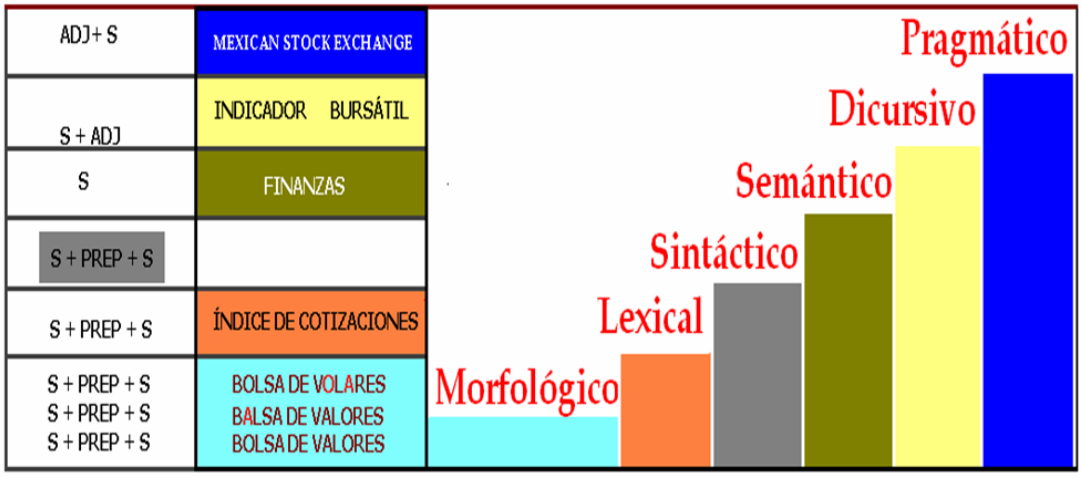
No obstante, si nuestro inglés universitario es suficientemente “bueno” para obtener la información que deseamos, y contamos con bibliotecarios comprometidos que nos facilitan las herramientas y técnicas esenciales y, por su parte, los SRI ofrecidos por lso proveedores de las BDI son “efectivos”, así lo dicen, ya que se emplean herramientas diversas y eficientes para la RI, ¿por qué aún enfrentamos altos niveles de frustración al tratar de recuperar la información que necesitamos?

Una de las grietas más conocidas por todos es la del idioma. ¿Por qué representa una grieta? En primer lugar, el idioma inglés no es nuestra lengua materna, pero tenemos que emplearla para obtener información para nuestro quehacer académico y es precisamente la falta del manejo apropiado lo que ocasiona desviaciones semánticas. En segundo lugar, la falta de herramientas lingüísticas como los diccionarios especializados y los tesauros, cuyo manejo y variedad del vocabulario controlado, en idioma inglés, permite acercarnos a una disponibilidad léxica ad-hoc a las fuentes consultadas.

Por ejemplo, en los seis niveles de comprensión de idioma se observan las ponderaciones conducentes al éxito o fracaso de nuestra búsqueda de información. Tomemos como caso que un usuario está interesado en la búsqueda de bolsa de valores, ilustrado en un breve análisis lingüístico de seis niveles, iniciado en español y finalizado, inevitablemente en inglés :

Buscando información sobre *bolsa mexicana de valores* (Mexican bag of values)

Niveles de comprensión de lenguaje



En el nivel morfológico se presentan algunas posibilidades de incurrir en un error semántico por la distribución de los grafemas, considerando el lenguaje natural; en el nivel léxico se presenta, el cambio ortográfico por la naturaleza del vocabulario controlado; el nivel sintáctico presenta el orden de los componentes, un sustantivo, una preposición, un sustantivo; en el nivel semántico, se observa la reducción de tres elementos a sólo uno, y el cambio de referente; en el nivel discursivo se da un cambio en el número de elementos y de referente, compuesto por dos elementos; finalmente, en el nivel pragmático, se toma como base que, si se trata de finanzas y la base de datos recupera principalmente información en inglés, se deberá modificar la sintaxis y el escribir con un vocabulario controlado.

Hay que observar que, a pesar de tratarse de dos términos, ambos funcionan como un sólo sustantivo, pero con la sintaxis indicada, primero stock, después exchange. En un caso extremo, el usuario podría sugerir, con base en un mal manejo de la lengua inglesa y de un diccionario español/inglés: “bag of values”, porque no se puede negar que bag significa bolsa y values, valores.

No obstante lo anterior, ¿por qué debemos recuperar información de las bases de datos en inglés? Algunas razones al respecto las desglosa Dungworth (1978:1), “el idioma inglés es el más empleado en el mundo por tres razones 1) porque ocupa la segunda posición en el rango mundial y sólo es superado por los hablantes de Chino Mandarín, 2) porque es el más empleado en el control de tráfico aéreo, almacenamiento y recuperación de información y 3) está mucho más difundido que cualquier otro y está considerado, según el reporte del Consejo Británico (1974-1975) como el idioma líder para la comunicación internacional”.

Un reporte de 2006 indica que el idioma inglés es la lengua franca de la ciencia y la tecnología, de la experimentación y los descubrimientos y su influencia ha sido tal que: Los journals en muchos países han cambiado de lengua, desde la II Guerra Mundial, de editarse en su lengua vernácula a publicarse en inglés. Gibbs (1995) describe cómo el journal medico mexicano Archivos de Investigación Médica mudó al inglés: primero se publicaron los resúmenes en inglés, después se ofrecieron traducciones al inglés de todos los artículos, finalmente se contrató a un editor norteamericano, y únicamente se aceptaron artículos en inglés y cambió su nombre a Archives of Medical Research. Este cambio de idioma es muy común en otros lugares (Graddol, D. 1997: 9).

Consecuencia de lo anterior es la tendencia de los científicos a buscar y recuperar la información que requieren las bases de datos que nos ocupan. Actualmente, la UNAM ofrece a su comunidad más del 85% de este tipo de bases, de las que en el año 2006, reporta la Biblioteca Digital de la UNAM, se hicieron 55,666 búsquedas de información en las diez bases de datos más consultadas. Nueve de ellas están dispuestas para su consulta en idioma inglés y el total de sus consultas fue de 54,533, que representa el 98%; la última base de datos, dispuesta en español, tuvo 1,133 consultas, lo que representa el 2%. Por su parte, la Internet Society ha reportado sus más recientes hallazgos en una encuesta sobre el idioma de las páginas web que hay en el mundo, siendo el 84.3% en inglés y el 1.2% en español.

Estadísticas de uso de bases de datos en BiDi-UNAM

Consultas	Bases de Datos más consultadas en 2006
25,723	Academic Search Premier
7,753	Elsevier Science
5,472	PsycINFO
4,168	WorldCat
3,183	ProQuest Psychology Journals
2,927	Science Citation Index
2,092	MEDLINE
2,057	Journals@Ovid Full Text
1,158	EJS (Ebsco Host Electronic Journal Service)
1,133	INFOLATINA

Lo anterior nos hace pensar en la necesidad de planeación del idioma para el cambio social, de acuerdo con Cooper (1989:182), “el idioma inglés conduce a reducir la grieta entre las variedades hablada y escrita y para incrementar el acceso a la educación formal, además, la planeación lingüística es una herramienta al servicio de tantas y

diferentes metas como la modernización económica, integración nacional, liberación nacional, igualdad social, racial, y el mantenimiento de las elites y su reemplazo por nuevas". Hasta aquí, las consideraciones del idioma. Pasemos a la revisión de la literatura.

Revisión de la literatura en TESIUNAM

Se efectuó una revisión de la literatura de tesis en el catálogo TESIUNAM considerando diferentes términos relativos a la recuperación de información y lingüística – en título; y con los términos bibliotecología y lingüística en carrera.

Se seleccionaron aquellas tesis cuyos títulos indicaran una relación estrecha con el propósito de esta investigación y, posteriormente, se revisaron las tablas de contenido de 25 de ellas, y se leyeron los capítulos relacionados con el idioma inglés para extraer la información que comprobara una línea de investigación al respecto. Sólo se obtuvo el siguiente párrafo relativo al interés de esta investigación (Balboa, A. 2005:115) "En este estudio se ve claramente que ya es bastante difícil para bibliotecarios y usuarios el entenderse en español; luego entonces, una barrera de idioma, dificulta mucho más la comunicación y es más frustrante". Por cuestiones de espacio, sólo se hace referencia a la tesis de la que se tomó el párrafo referido.
El resto de los resultados fue el siguiente:

Esfuerzos documentados en tesis

Palabras / campos:	# Registros
Recuperación(en Tit). Lingüistica (en carrera)	SEIS
Lingüísticas (en Tit). bibliotecología (en carrera)	CERO
Lingüística (en Tit). bibliotecología (en carrera)	UNO
Búsqueda recuperación información (en Tit).	CUATRO
CD ROM (en Tit). bibliotecología (en carrera)	DOS
Bases de datos (en Tit). bibliotecología (en carrera)	OCHO
Usuarios (en Tit). bibliotecología (en carrera)	TREINTA Y UNO
Usuario (en Tit). bibliotecología (en carrera)	DOS
Habilidades informativas (en Tit)	UNO

Por otro lado, también se realizó una búsqueda en la bases de datos LISA, en –línea con el siguiente resultado. Cabe mencionar que, al igual que en las tesis, sólo se hará referencia a los registros de los que se tomó una referencia sólida:

En la base de datos LISA

Términos empleados	Número de registros	Notas
information retrieval	22362	Se restringió la búsqueda agregando "Spanish"
(information retrieval) and Spanish	240	Se seleccionaron 3 registros de 22, que abordan el idioma español con propósitos adecuados a este estudio (5, 10 y 12)
(information retrieval) and Spanish and linguistics	1	Este registro presenta las palabras distribuidas a lo largo del registro.
(information retrieval) and Spanish and deconstruction	0	
TI=(information retrieval Spanish)	0	

La perspectiva deconstructiva

La aparente facilidad de recuperar información de un texto (artículo, libro, reporte, etc.) en bases de datos sin ningún otro esfuerzo que ingresar los términos necesarios en el teclado de una computadora en que reside un gran cantidad de datos, contrasta con el nivel de frustración que los usuarios experimentamos al comprobar que no es así. Aparentemente, la intervención humana todavía se requiere de manera física, crítica y reflexiva.

Una perspectiva deconstructiva podría ayudar a la creación de un peldaño que ayude, en la medida de lo posible, a acortar la distancia entre la satisfacción y el infortunio, por medio de la irrupción de deconstruir el sistema de oposiciones conceptuales como son lo literal vs lo metafórico, el habla vs la escritura, lo inteligible vs lo sensible,

Es pertinente la propuesta de un enfoque que permita explorar las tensiones y contradicciones no sólo del (o los) textos contenidos en una BD y su significado, sino de las relaciones lingüísticas que se realizan en la interacción entre los usuarios humanos y la máquina con que se opera la BD y, necesariamente también, su significado.

De manera sencilla, si partimos de la siguiente hipótesis: Si poseo buen manejo del idioma inglés, a pesar de no ser mi lengua materna, si me asiste un bibliotecario profesional y consulto una BD apropiada a mis necesidades, encontraré el (los) texto(s) con la información que necesito.

Más del 50% de usuarios estaría en desacuerdo con la hipótesis anterior porque sencillamente han experimentado lo anterior con resultados desafortunados. En este

punto es pertinente comentar aún para los nativo hablantes del idioma inglés, la experiencia de frustración a causa de la ambigüedad en la búsqueda y recuperación de información se hace presente en el momento de revisar los resultados arrojados: (Jaffe, J. 1988b:759) “Los usuarios quedaban confundidos con las citas sin relevancia aparecidas en las búsquedas booleanas, aunque todos los elementos de las mismas estaban representados en las citas.” Esta afirmación se hizo después que estudiantes del Sweet Briar Collage, en Virginia, consultaron una base de datos. Lo anterior revela la existencia de una grieta, de un malentendido que se puede interpretar como “ la imposibilidad de aislar un sentido originario principal en el centro de una construcción conceptual o el conjunto de una obra” (Peñalver, P. 1989:15).
.

En la actualidad, sabemos que las técnicas genéricas empleadas para RI son, a grosso modo, dos: la comparación de las palabras de la búsqueda contra el índice de la BD de que se trate y la de atravesar la BD con la ayuda de vínculos de hipertexto o de hipermedios, por medio de operadores booleanos. Asimismo, el inevitable uso de lenguaje natural y el vocabulario controlado, son aún insuficientes para evitar la ambigüedad que conlleva el lenguaje humano para obtener la información deseada en una BD, en una clara oposición de la inteligencia artificial y la sensibilidad humana.

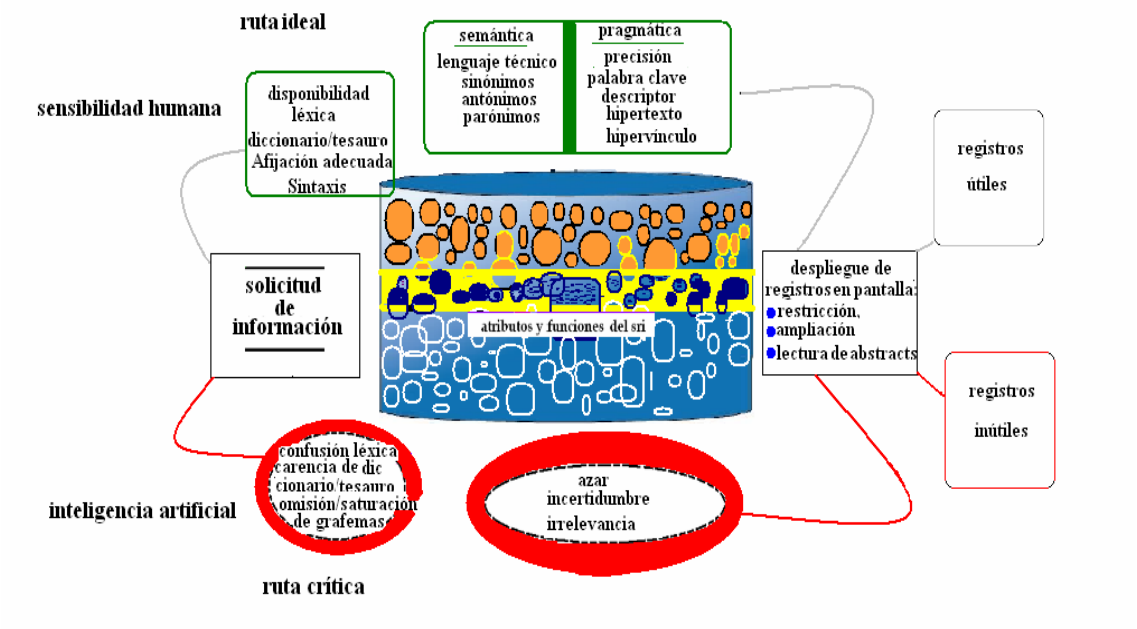
Este escenario sugiere considerar las limitantes de los sistemas de recuperación para reconocer los significados de las imágenes, idiomas o conceptos en la experiencia deconstructiva de Jeff Bezzos, ejecutivo en jefe de Amazon.com, quien creó un servicio en-línea que incluye a humanos, denominado Amazon Mechanical Turk, fenómeno que en sus palabras describe de la siguiente manera: “Normalmente, un humano hace una petición a una computadora, y la computadora hace su tarea. Pero las inteligencias artificiales artificiales como la Mechanical Turk invierten todo eso. La computadora tiene una misión que es fácil para el humano, pero extraordinariamente difícil para la computadora. De este modo, en vez de solicitar el servicio de computadora para realizar la función, se pide a un humano.” (Pontin, J. 2007: s/p)

Los diversos esfuerzos existentes para obtener la información deseada que van desde realizar procesos sometidos al azar y que son objeto de análisis estadísticos hasta los análisis gramaticales empleados en la minería de datos, hace que los usuarios finales estemos más esperanzados a obtener menos grados de frustración que los actuales. No obstante, dado que el estudio, los descubrimientos y las aplicaciones tecnológicas en los SRI no se detienen para indexar documentos académicos con formatos diferentes a los de puro texto, como imagen, vídeo, audio y sus combinaciones, nos reduce la esperanza de obtener información con mayor facilidad.

La solución pragmática

Tomando en cuenta los argumentos anteriores, ya se está en condiciones de sugerir una solución pragmática con base en una serie de operaciones para resolver el problema de la incertidumbre: el algoritmo lingüístico, marcado con los cuadros verdes - en el orden descrito - y que se apoya en el plano de la sensibilidad humana, en oposición a la

inteligencia artificial, desde nuestra perspectiva deconstructiva, acusada líneas atrás.



El punto de partida es la solicitud de información con dos rutas a seguir: a) la ruta ideal, representada en la parte superior del esquema y b) la ruta crítica es la acostumbrada por los NHI y por los NHE. El cuadro del centro representa el universo de datos contenidos en el SRI; los círculos del interior destacados en color naranja representan datos relevantes recuperados por medio del empleo de los cuadros superiores distinguidos en color verde; el área destacada en amarillo, son aquellos datos que representan potencialidad de relevancia, pero que requieren de un empleo más refinado en el manejo de vocabulario controlado; las figuras destacadas en color azul representan datos de naturaleza entrópica, ya que, ortográficamente serían afines, pero semánticamente resultarían fragmentados por alguno(s) niveles de comprensión de lengua ilustrados anteriormente o por desemejanza en la relación SIGNIFICADO vs SIGNIFICANTE, considerando, asimismo, la interferencia del idioma español y su realización en idioma inglés: *bolsa/bag; de/of; valores/values ## stock exchange.*

y depende, en gran medida, de la inteligencia artificial dispuesta en los SRI de las BDI y que difícilmente resuelve los problemas de incertidumbre, a saber (Cfr. Allwood, J. 1995: 29)

I) Si no se encuentra información…
es porque se carece de disponibilidad léxica
o porque está mal el orden de las palabras
o porque el (los) término(s) empleado(s) no es (son) equivalente(s) en significado
o porque el contexto no es el apropiado
o porque no existe en la base de datos que se esté empleando

II) Si la información no es la deseada
es porque el lenguaje técnico no es el apropiado

o porque la ortografía no es la correcta
o porque la palabra clave/descriptor no es tal

En cuanto al espacio de la inteligencia artificial, es de esperar que los SRI de las BDI, incluyan, puntos de recuperación que incluyan campos referentes a imágenes, audio, vídeo y sus posibles combinaciones.

Conclusiones

A lo largo de este trabajo se planteó la posibilidad de resolver la problemática de la RI en BDI, con base en una perspectiva deconstructiva lingüística, y centrada en NHE poniendo en evidencia las variedades existentes en seis niveles para la comprensión del idioma inglés en cuanto a los grafemas y a la sintaxis. Sin este matiz, difícilmente se podría entender la existencia de ángulos diferentes a los ya conocidos, aunque no sean lo únicos.

Se tomaron en cuenta tres aproximaciones a la resolución de la frustración que experimentan los usuarios: las realizadas por los tesistas egresados de la UNAM, en sus trabajos recepcionales, registrados durante los últimos años en el catálogo TESIUNAM; una revisión de la literatura que ofrece la base de datos LISA; y algunas consideraciones sobre las herramientas empleadas en los SRI de las BDI con el mismo resultado: insuficientes.

Asimismo, se ilustraron algunas circunstancias históricas y tecnológicas relativas a la influencia de la lengua franca de nuestros tiempos y que sirven para entender las tendencias y necesidades implícitas para los no NHI. Se describieron algunas actitudes de los involucrados en el manejo de las BDI y las interrogantes que surgen durante su quehacer académico. Se quiera o no, estamos supeditados a acercarnos, invariablemente, al manejo adecuado del idioma inglés.

Finalmente, se describió un modelo que incluye un algoritmo lingüístico, y dos oposiciones presentes durante la ruta de búsqueda y recuperación de información, a saber, la sensibilidad humana y la inteligencia artificial, ponderando la ruta crítica, la ruta ideal y contrastando los atributos humanos y los de los SRI de las BDI. Se puede concluir que estimando la ruta ideal y las condiciones de incertidumbre, se podrán lograr mayores satisfacciones si únicamente recurrimos a lo que ofrecen los proveedores de las BDI.

Bibliografía

ALLWOOD, Jens. Dialog with a cooperative information system. Sweden: Dept. of Linguistics. University of Göteburg. 1995.

DERRIDA, J. La deconstrucción en las fronteras de la filosofía : la retirada de la metáfora / Traducción de Patricio Peñalver. Universitat Autònoma de Barcelona: 2001.122 p.

DUNGWORTH, D. The future of English as a World Language. Lebende Sprachen. 1978. Vol. 23, no. I, p. 1-3.

ELY, Donald. ERIC: multinational/multidisciplinary. Revista Interamericana de Bibliotecología. 1985; Vol. 8, no. 2. July-Dec. p. 49-61.

GRADDOL, David. The Future of English?: A guide to forecasting the popularity of the English language in the 21st century. The British Council. [En línea] <http://www.britishcouncil.org/learning-elt-future.pdf> [Consulta: 9 julio, 2007]

IBARRA Contreras, R. Aprovechamiento y optimización de los recursos tecnológicos en la búsqueda y recuperación de información en CD-ROM basados en estrategias lingüísticas. Tesis de maestría UACPyP – CELE UNAM, 1999. <http://pbidi.unam.mx/cgi-bin/ezpmysql.cgi?url=http://132.248.9.9:8080/tesdig/Procesados_1999/271153/Index.html> Acceso restringido al documento electrónico.

JAFFE, J. For undergrads: Infotrac MAGAZINE INDEX PLUS or WILSONDISC with Reader's Guide & Humanities Index. American Libraries, 1988. Oct. 1988 pp.759-61.

LIDDY, Elizabeth D. Whither Come the Words? Paper presented at the CENDI Subject Analysis and Retrieval Working Group Conference "Controlled Vocabulary and the Internet," September 29, [Presentación power point]. [En línea] <http://cendi.dtic.mil/presentations/liddy.PPT> [Consulta: 5 julio, 2007].

PONTIN, J. Artificial Intelligence, With Help From the Humans. <http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html?ex=1184731200&en=8d71530b581d5e00&ei=5070> [Consulta: 25 marzo, 2007].