ANÁLISIS DE LOS RECURSOS LINGÜÍSTICOS UTILIZADOS EN LOS SISTEMAS MULTILINGÜES DE BÚSQUEDA DE RESPUESTAS

Juncal Gutiérrez-Artacho[1] y María-Dolores Olvera-Lobo[2]

[1] Universidad de Granada, España

[2] CSIC, Unidad Asociada Grupo SCImago, Madrid y Universidad de Granada, España

1. Introducción

En el ámbito de la recuperación de información (en adelante, RI) se están creando herramientas documentales e informáticas (monolingües y multilingües) que pueden ayudar substancialmente a los especialistas en su trabajo −además de resultar útiles para otros usuarios con necesidades de información de lo más diversas−. El desarrollo de las herramientas multilingües se encuentra todavía en evolución y necesita varios años de estudios e investigación para su mejora y aplicación. Uno de los principales problemas a los que se enfrentan estas herramientas es la traducción (Diekema, 2003), tanto de las consultas planteadas por los usuarios como de las fuentes documentales que responden a las mismas. Por tanto, ante el creciente auge en la investigación, desarrollo y creación de sistemas multilingües de RI, consideramos necesario realizar un estudio que se centre en el análisis y evaluación de los recursos utilizados por un tipo de estos sistemas como son los sistemas multilingües de búsqueda de respuestas (en adelante, BR).

Aunque las investigaciones en esta área se iniciaron hace algo más de una década, estos sistemas son unos auténticos desconocidos fuera del ámbito de la RI. Realizar un estudio desde la disciplina de la traducción podría ofrecer una perspectiva distinta de la

problemática de la traducción y de sus recursos en los sistemas multilingües de BR. Los actuales investigadores en el área intentan buscar nuevos métodos para que la RI sea lo más eficiente posible pero sin atender detenidamente a los problemas lingüísticos. Sin embargo, si no se encuentra una solución óptima en relación a la traducción y a los recursos utilizados, difícilmente el sistema podrá recuperar una información relevante para el usuario. Por este motivo, la traducción cobra un protagonismo fundamental en este entorno y permite analizar el problema desde una nueva perspectiva.

Esta investigación se plantea como primer objetivo general el análisis e incorporación de la disciplina de la traducción en el estudio de los sistemas multilingües de BR. Nuestro segundo objetivo general es proceder al análisis y evaluación de los recursos y herramientas lingüísticos utilizados en estos sistemas. Estos objetivos confieren una nueva perspectiva al problema de la recuperación de información multilingüe. Además, como objetivos específicos la identificación de los principales tipos de herramientas y recursos lingüísticos útiles en los procesos de RI multilingüe, concretamente en el caso de los sistemas multilingües de BR, y el establecimiento del grado de utilización real que hacen los sistemas multilingües de BR de cada uno de los recursos y herramientas analizados.

2. Estado del arte

En el entorno de la Web la sobrecarga de información se deja sentir aún más que en otros contextos. De esta forma, en demasiadas ocasiones, al plantear una determinada consulta en las herramientas de búsqueda de información web (buscadores, directorios o metabuscadores) el número de páginas web recuperadas resulta excesivo y no todas

ellas son relevantes ni útiles para los objetivos del usuario. Por ello, los profesionales de diversos ámbitos comienzan a reconocer la utilidad de otros tipos de sistemas, como los sistemas de BR, como método para la obtención de información especializada de forma rápida y efectiva (Crouch et al., 2005; Lee et al., 2006).

Tradicionalmente, la RI se ha entendido como el proceso, totalmente automático, en el que dada una consulta (que, supuestamente, expresa la necesidad de información del usuario) y una colección de documentos, el sistema devuelve una lista ordenada de documentos potencialmente relevantes para esa consulta. Un sistema de RI con funcionamiento óptimo recuperaría todos los documentos concurrentes (lo que implica una cobertura completa) y sólo aquellos documentos que son relevantes (precisión perfecta). Este modelo tradicional lleva consigo muchas restricciones implícitas como: a) la suposición de que los usuarios del sistema buscan documentos (textos completos), no respuestas, y que son los documentos, como tales, los que responden y satisfacen una consulta; b) que el proceso debe ser directo y unidireccional en lugar de interactivo; c) y por último, que la consulta y el documento están escritos en la misma lengua.

La RI multilingüe implica, al menos, la participación de dos lenguas en este proceso. En un entorno multilingüe como es el de la Web, la mayoría de los sistemas de RI tienen la limitación de encontrar documentos sólo en el idioma en que se escribe la consulta o bien incorporan sistemas de traducción automática, que únicamente resultan útiles cuando los documentos ya han sido localizados, pero no facilitan un medio efectivo para salvar la barrera del idioma en el proceso de búsqueda.

Un paso en la evolución hacia la mejora de la RI son los sistemas de BR. Se presentan como una alternativa a los tradicionales sistemas de RI tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de presentar al

usuario una lista de documentos relacionados con la búsqueda (Jackson & Schilder, 2005), de modo que el usuario no ha de leer documentos completos para obtener la información requerida.

El funcionamiento de los sistemas de BR se basa en los modelos de respuestas cortas (Blair-Goldensohn, 2004), ya que se divide la pregunta asignando a la palabra clave una etiqueta que indica el tipo de preguntas que puede responder. El sistema reemplaza esa etiqueta por las palabras adecuadas para poner a disposición de los usuarios una selección de textos que responden correctamente a la consulta (Pérez-Coutiño et al., 2004). La ventaja principal es que el usuario no ha de leer documentos completos para obtener la información requerida puesto que el sistema ofrece la respuesta correcta en forma de un número, un sustantivo, una frase corta o un fragmento breve de texto.

Aunque existen diferentes patrones a la hora de plantear las preguntas en los sistemas de BR, la mayoría se caracterizan por aceptar preguntas expresadas a través de partículas interrogativas (qué, cómo, quién, por qué, cuándo, dónde), o expresadas mediante una forma imperativa. Planteada la pregunta en el motor de búsqueda del sistema, se procede a analizar la pregunta separando la palabra o palabras claves, luego se localiza y extrae una respuesta a partir de diferentes fuentes –dependiendo de la especialización del sistema se utilizarán unas u otras fuentes (Olvera-Lobo & Gutiérrez-Artacho, en prensa)–, y finalmente, se evalúa y elimina aquella información redundante o que no responde correctamente a la pregunta planteada para, posteriormente, elaborar y presentar una o varias respuestas concretas que supuestamente satisfacen la necesidad del usuario (Cui et al., 2004; Tsur, 2003).

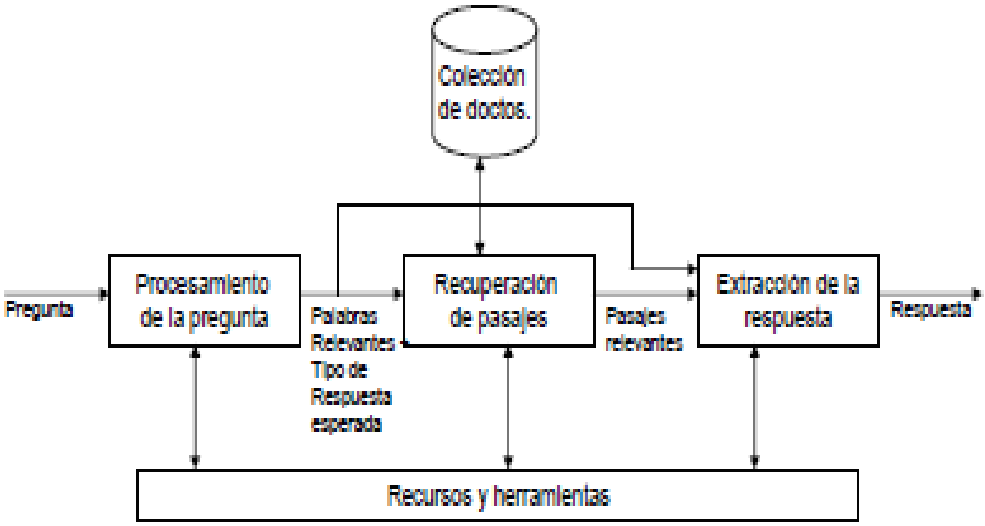

Figura 1. Representación de un sistema clásico de BR monolingüe (Aceves 2008)

Estos sistemas suelen tener una sencilla interfaz con un motor de búsqueda en el que los usuarios plantean su pregunta, algunos de ellos facilitan la lista de las últimas cuestiones introducidas para facilitar al usuario la compresión acerca del funcionamiento del mismo. Para el tratamiento y gestión de las preguntas, los sistemas de BR aplican algoritmos y métodos del análisis lingüístico y del procesamiento del lenguaje natural con el fin de identificar sus componentes y determinar el tipo de respuesta esperada (Zweigenbaum, 2005). Este análisis consiste normalmente en utilizar una variedad de tipos de preguntas estándar en los que se reemplazan ciertas palabras por las etiquetas aceptadas por el sistema.

Los sistemas de BR pueden ser de dominio general –si puede atender consultas de temas muy diversos, como START[1] o NSIR[2]– o de dominio específico si se centran en un ámbito determinado, como MedQA[3] (Frank et al., 2006). Éstos son algo más

[1] http://start.csail.mit.edu/ (Última consulta el 8 de Enero del 2010)
[2] http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi (Última consulta el 8 de Enero del 2010)
[3] http://monkey.ims.uwm.edu:8080/MedQA/ (Última consulta el 8 de Enero del 2010)

frecuentes debido a que permiten el uso de recursos lingüísticos especializados con lo que se consigue una mejor precisión en las respuestas ofrecidas.

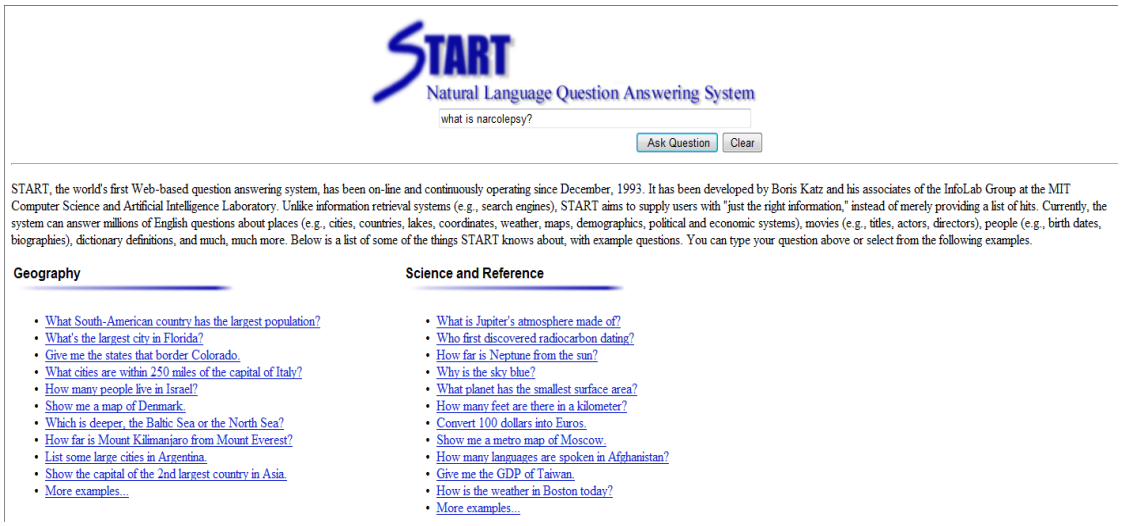

Figura 2. Motor de búsqueda en el sistema de búsqueda de respuestas START

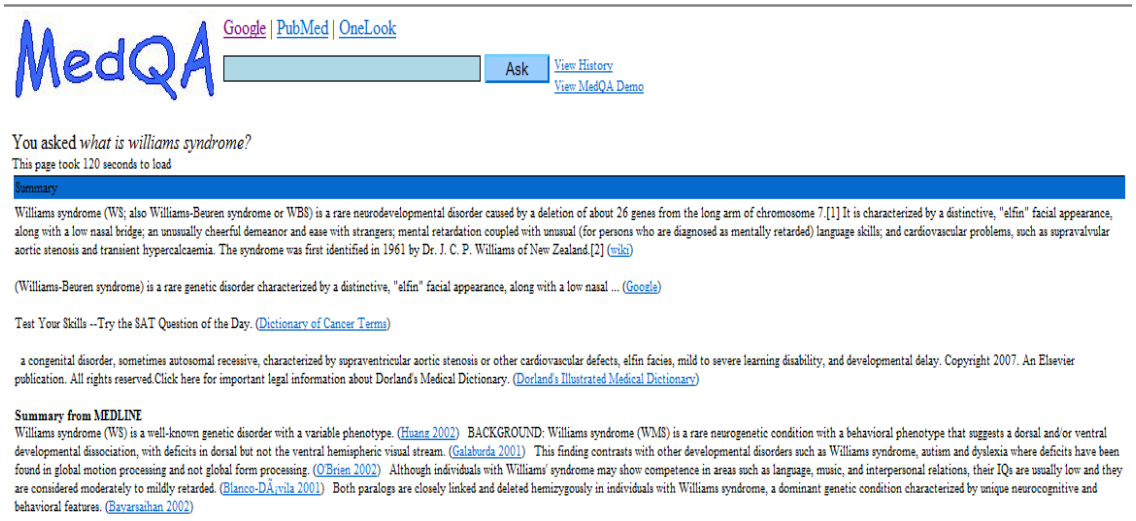

Figura 3. Respuestas ofrecidas por el sistemas MedQA para la pregunta *What is William syndrome?*

Otro de los aspectos clave de estos sistemas es que la relación sistema-usuario no es unidireccional y se establece una interacción con el mismo que ayuda al sistema de BR a encontrar mejores respuestas, y a su vez, el sistema ayuda al usuario a encontrar la respuesta más rápidamente. No obstante, todavía es necesario profundizar en el diseño de estos sistemas interactivos que hagan posible la existencia de un verdadero *feedback*

entre preguntas y respuestas, y que el usuario se comunique a nivel conversacional con el sistema.

A pesar del avance que supone el poder contar con herramientas de búsqueda de información de este tipo, los sistemas de BR presentan algunas restricciones. En primer lugar, muchos de los sistemas han sido desarrollados únicamente como prototipos, o bien están disponibles únicamente como *demos*, y sólo en casos muy poco frecuentes se han comercializado. Algunos investigadores diseñan y crean sistemas que se presentan y discuten en diferentes foros y congresos pero, bien porque su utilidad se limita a contextos muy concretos, bien por sus dificultades de implementación, y salvo contadas excepciones, finalmente no se desarrollan para los usuarios finales.

Dentro de los sistemas de BR, debemos detenernos en los vinculados a la RI multilingüe, los cuáles implican, al menos, dos lenguas, y permiten plantear las consultas en varios idiomas y recuperar información en todas las lenguas aceptadas por el sistema (Diekema, 2003). Estos sistemas son capaces de operar en una colección de documentos multilingües dada una consulta determinada, y recuperar aquella información relevante que responda a la misma, independientemente del idioma utilizado al plantear la consulta (Grefenstette, 1998). Dentro de la RI multilingüe, debemos destacar el caso especial del objeto de nuestro estudio, los sistemas multilingües de BR, que son aquellos sistemas donde el idioma en el que se plantea la pregunta puede ser diferente a la lengua en la que está escrito el documento recuperado, pero se diferencian del resto de sistemas de la RI multilingües en que éstos no recuperan documentos completos sino que responden con una respuesta corta a la consulta planteada. Normalmente el funcionamiento de estos sistemas es muy similar a los de

BR monolingües, solamente se incorpora el módulo de la traducción y/o la herramienta o recurso lingüístico que llevará a cabo la recuperación translingüe (véase figura 4).

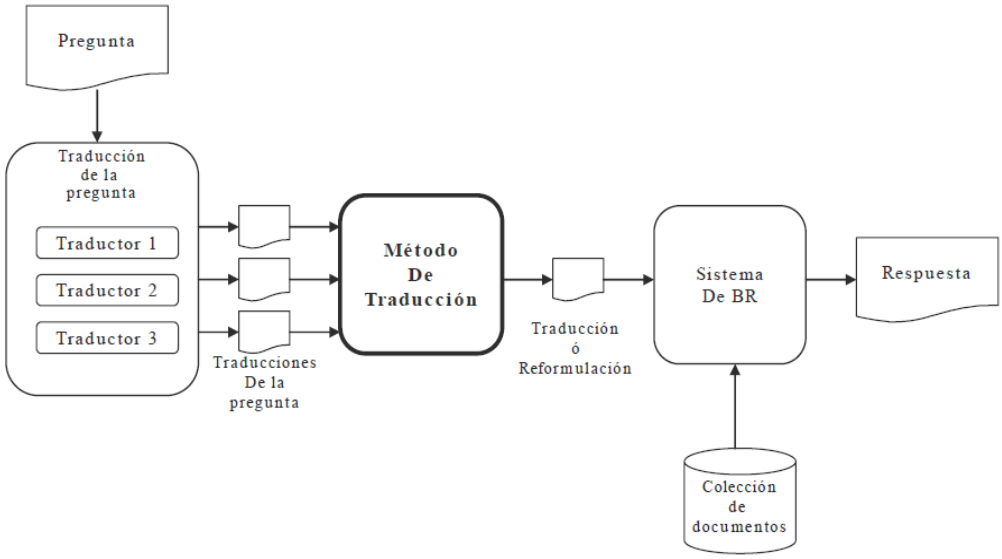

Figura 4. Representación de un sistema multilingüe de BR de traducción de preguntas

Una disciplina que ocupa un lugar relevante en estos sistemas es la traducción ya que las consultas y los documentos no siempre comparten el mismo idioma. Los principales problemas traductológicos identificados hasta el momento son la ambigüedad léxica, la falta de cobertura traductora, los lexemas multimodales y los errores en los recursos léxicos (Diekema, 2003). Sin embargo, todavía no se ha realizado ningún trabajo que sitúe a la disciplina de la traducción al nivel correspondiente que debiera ocupar en estos sistemas.

3. Metodología

Se ha adoptado una metodología empírica experimental en donde se estudian y recogen datos acerca de cada una de las diferentes herramientas y recursos lingüísticos utilizados por estos sistemas, así como de su uso e implementación.

Nuestro primer paso fue la identificación de los principales congresos, conferencias y foros que han tratado y tratan los sistemas multilingües de BR, con la intención de poder identificar, analizar y comparar los diferentes tipos de recursos lingüísticos utilizados. Aunque cada año se celebran un número mayor de congresos centrados en la RI, no todos ellos contienen un apartado dedicado exclusivamente a la investigación sobre los sistemas de BR, y aún menos los que tratan el problema multilingüe. Sin embargo, hemos identificado varios congresos o foros donde la investigación de los sistemas multilingües de BR ocupa un lugar primordial en su celebración.

En total hemos analizado ciento sesenta y cinco publicaciones que se han presentado desde el año 2000 hasta el 2008 en las diferentes conferencias. Los años que más publicaciones sobre sistemas multilingües de BR han sido el 2005 y el 2008. Ha habido una progresión ascendente en el interés de las publicaciones, siendo su año auge el 2008. Sin embargo, en el año 2007 disminuye el interés porque se apostó por el estudio de diferentes tipos de sistemas de BR como los de imagen, voz y dominios de especialización.

Nuestro trabajo ha tenido una segunda fase en donde se han estudiado los recursos utilizados por los sistemas multilingües de BR existentes. Para ello, se han identificado y analizado todos los sistemas, tanto monolingües como multilingües, disponibles para usuarios finales. En algunos sistemas de BR, ha sido relativamente sencillo obtener información sobre el recurso lingüístico utilizado ya que son de libre acceso y sus

desarrolladores facilitan todas las publicaciones que han escrito sobre el sistema. Sin embargo, estos son solamente un número reducido, puesto que la mayoría son prototipos y no están totalmente desarrollados, es decir, no se puede acceder a ellos o no funcionan correctamente. Por ello, ha sido tan importante realizar la fase de observación documental mediante el análisis de las publicaciones ya que los desarrolladores publican todos sus avances en los foros.

4. Resultados y discusión

Gracias al análisis de las publicaciones, se pueden encontrar cinco tipos de recursos lingüísticos principales utilizados por los sistemas multilingües de BR, a saber, las bases de datos, los corpora, los diccionarios, las ontologías y los tesauros, así como dos tipos de herramientas lingüísticas usadas por estos sistemas –los traductores automáticos y las gramáticas computacionales–. Estos recursos y herramientas, junto con sus diferentes tipos y subtipos, no funcionan de igual modo y usan diferentes formas de procesar la información. En ocasiones, utilizar un único recurso no es suficiente y se debe hacer uso de varios de ellos para conseguir mejores resultados. Aunque anteriores trabajos realizados (Diekema, 2003) han identificado cuatro principales fuentes de traducción aplicadas a CLIR –ontologías, diccionarios bilingües, traductores automáticos y corpora, hemos comprobado que esta tipología ha aumentado en los últimos años y ciertos recursos han comenzado a ser bastante utilizados.

Las investigaciones y avances obtenidos en los sistemas multilingües de BR en los últimos años se refieren principalmente a la incorporación más efectiva de nuevos recursos, a la creación de sistemas más rápidos y eficientes, a la obtención de una mayor transparencia en los resultados, entre otros. Sin embargo, todavía hay un reto que no se

ha solucionado completamente: la traducción. La traducción se debería tratar en cada uno de los sistemas, independientemente del recurso o herramienta que se utilice. Los sistemas ya existentes han probado distintos modos de acercarse por medio del material que se traducía.

Hemos comprobado que el recurso que más se utiliza por los sistemas multilingües de BR son los traductores automáticos, seguidos de los corpora y diccionarios (véase figura 5). Según Nguyen et al. (2009), los tres recursos más utilizados hasta hace relativamente poco habían sido estos tres, por lo cual comprobamos que la situación no ha cambiado mucho. Aunque sí es verdad que ciertos recursos se están aumentando en popularidad en los últimos años.

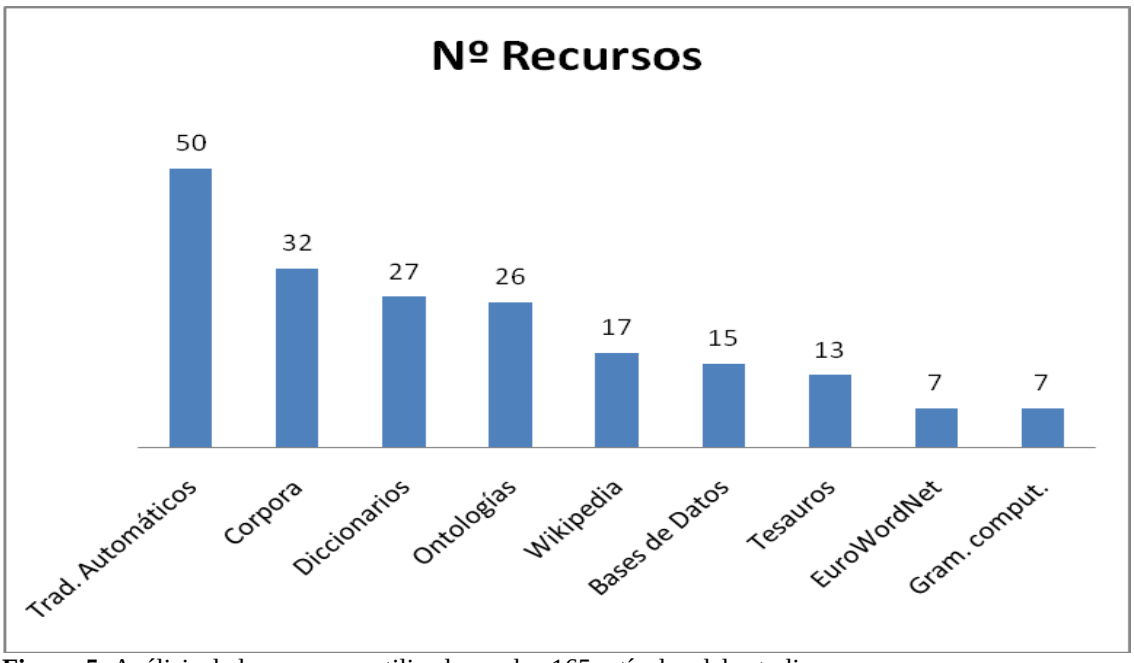

Figura 5. Análisis de los recursos utilizados en los 165 artículos del estudio

Los traductores automáticos han sido utilizados en 50 de los 165 artículos analizados. Esta herramienta se suele incorporar de manera individual o en combinación con otros recursos lingüísticos para proveer una cobertura mayor. Aunque la mayoría de los

autores confirman los problemas de ambigüedad y la mala calidad de los textos, prefieren usar esta herramienta porque es una de las más baratas y más fácil de incorporar a los sistemas. Esta herramienta suele dar mejores resultados en los sistemas multilingües de BR de dominio general que en los de dominio específico, ya que no son capaces de identificar ni traducir ciertos términos especializados correctamente.

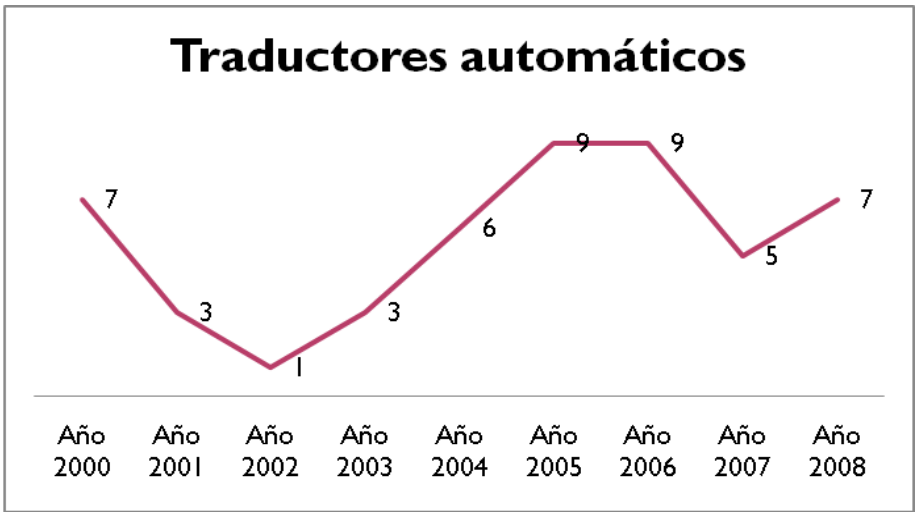

Figura 6. Análisis del uso de los traductores automáticos por años

No obstante, el uso de los traductores automáticos está disminuyendo en los últimos años. Hemos comprobado que al comienzo del estudio de estos sistemas los traductores automáticos fueron utilizados por 7 de los 9 artículos analizados ese año. Sin embargo, el número disminuye sustancialmente en los siguientes tres años (2001, 2002, 2003). A partir del 2006, aumenta el número de sistemas multilingües de BR que vuelven a usar esta herramienta, pero ya no es individualmente como en los primeros años sino en combinación o apoyando a otros recursos lingüísticos. Los últimos años se sigue usando pero en un menor número de sistemas.

El segundo recurso más utilizado son los corpora (32 veces), aunque si tenemos en cuenta que en este recurso hemos incluido todas las variantes del mismo, se puede

comprender su posición. Lo que más sorprende de este recurso ha sido su evolución casi sistemática en los últimos años y su auge en el año 2008, ya que 11 de las 30 publicaciones analizadas utilizaron ese recurso (véase figura 7). En el año 2007, vemos que hay un descenso considerable de su uso pero se puede atribuir en parte a que de ese año solo hemos considerado 20 publicaciones.

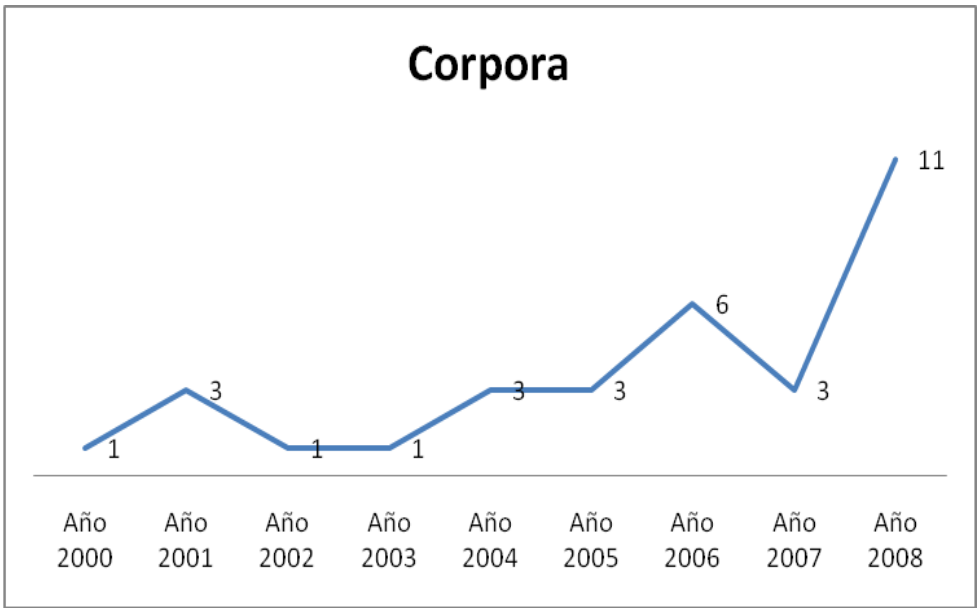

Figura 7. Análisis del uso de los corpora por años

Los corpora lingüísticos son unos recursos muy útiles para los sistemas de dominio especializado, ya que si se realiza una traducción del documento completo con traductores profesionales o revisados por ellos, la información que reciban los usuarios será completa y correcta. También es muy buena idea lo que han realizado algunos desarrolladores que han utilizado un corpus en la Web con páginas disponibles en varios idiomas, de modo que solventan los dos principales problemas que plantea lo anterior: el coste computacional y el almacenamiento.

El tercero y cuarto recurso más utilizado por estos sistemas son los diccionarios y las ontologías, con 27 y 26 apariciones respectivamente. Los diccionarios son junto con los

traductores automáticos y los corpora los recursos utilizados tradicionalmente por estos sistemas, de ahí que se compruebe una evolución parecida entre los tres recursos (véase figura 8). Sin embargo, los problemas gramaticales y de ambigüedad han disminuido su uso, de modo que solamente 5 de los 79 sistemas estudiados en los últimos cuatro años usan este recurso lingüístico.

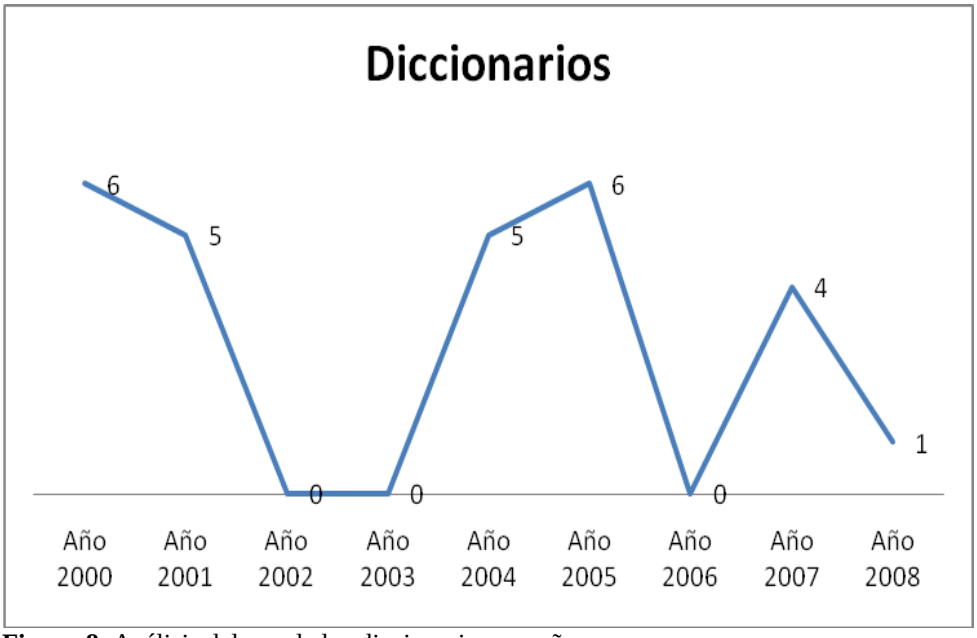

Figura 8. Análisis del uso de los diccionarios por años

Totalmente al contrario le sucede a las ontologías (véase figura 9). Aunque en los primeros años este recurso no era utilizado, a partir del año 2004 comienza su repunte siendo incorporado paulatinamente a los sistemas multilingües de BR. Como ya hemos comentado, las ventajas de las ontologías son muchas y sobre todo, en los sistemas de dominio especializado. Muchos de los sistemas se componen de textos que han sido traducidos completamente a las diferentes lenguas de trabajo, con lo que las relaciones son establecidas fácilmente. Otra ventaja es que ya hay ontologías multilingües y que muchos grupos de investigación trabajan minuciosamente en las diversas relaciones que se pueden establecer entre los términos, lo que asegura la correción en el producto final.

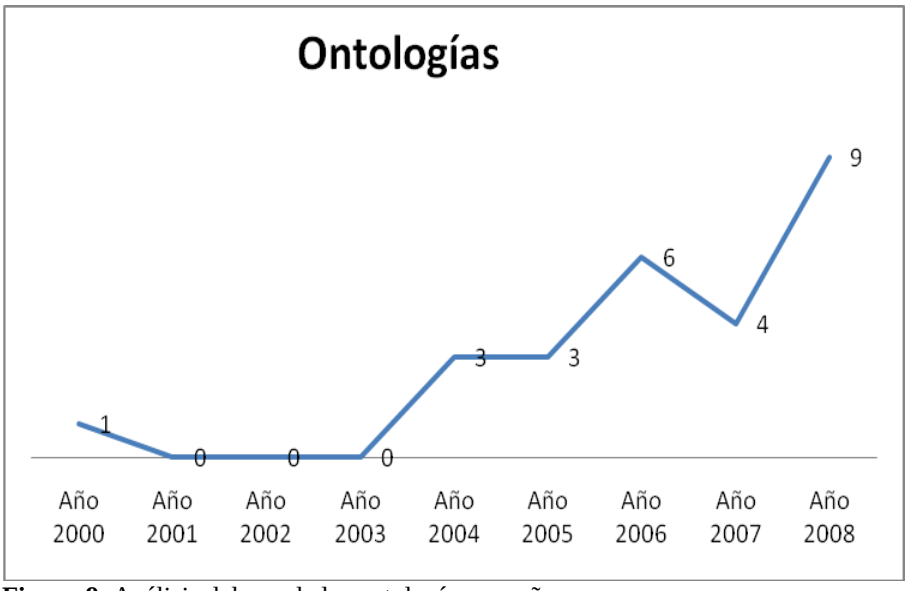

Figura 9. Análisis del uso de las ontologías por años

Además, hemos analizado el resto de recursos y herramientas utilizados por estos sistemas. Gracias a este análisis hemos comprobado el uso de cada uno de ellos, así como de sus características, ventajas y desventajas. Con todos estos datos, ya podemos decidir qué recursos son los más útiles y cuáles pueden ofrecer un producto final mejor. Además, este análisis nos ha dado una visión general de cómo se encuentra el problema de la traducción en los sistemas multilingües de BR y las posibles pautas que debieran seguirse en un futuro.

5. Conclusiones

El interés en la mejora de los sistemas de RI ha llevado a los investigadores a estudiar y desarrollar nuevos sistemas más complejos, ya que no recuperan textos completos que respondan a la necesidad del usuario sino que presentan exclusivamente la parte del texto que responde a su pregunta. Estos sistemas se denominan sistemas de búsqueda de respuesta y son el nuevo objetivo de los investigadores en RI. Desde hace más de una

década, Internet está sufriendo un cambio en su formato ya que cada vez hay disponibles más sitios web que no se encuentran exclusivamente en inglés. Los gestores web prefieren crear páginas disponibles en el idioma de los usuarios a los que se destine la información, o en varios idiomas atendiendo a las particularidades de cada cultura. Este hecho no ha pasado desapercibido entre los investigadores de los sistemas de BR puesto que han comenzado a desarrollar sistemas multilingües que son capaces de recuperar respuestas en más de dos idiomas.

Ante esta situación nace nuestro interés por estudiar la situación lingüística de los sistemas multilingües de BR hasta el momento con la intención de profundizar y buscar soluciones en trabajos futuros. Por ello, nuestro primer objetivo fue el estudio de las principales publicaciones en los últimos diez años, es decir, desde el año 2000 al 2008. En total, estudiamos 165 artículos que se habían presentado en las principales conferencias, y se extrajeron todos los datos posibles que pudieran darnos una visión general de la situación. Se estudiaron, definieron y analizaron los recursos y herramientas más utilizados: traductores automáticos, los corpora, diccionarios y las ontologías.

Los traductores automáticos siguen siendo la opción más utilizada a pesar de los problemas de ambigüedad reconocidos por los autores. El bajo coste computacional y las pocas gestiones necesarias lo convierten en la opción más recurrida por los desarrolladores. En nuestra opinión, esta herramienta no es totalmente inadecuada para la RI si es combinada con otros recursos o evaluada por especialistas lingüísticos –como los traductores. Hemos comprobado que los tres recursos tradicionalmente más populares (traductores automáticos, corpora y diccionarios) siguen manteniendo su posición, aunque algunos recursos están gradualmente aumentando el número de

investigadores adeptos a sus resultados. Todos estos datos permiten dilucidar que la evolución en el uso de éstos puede dar giros inesperados que deberán ser estudiados y evaluados en investigaciones futuras.

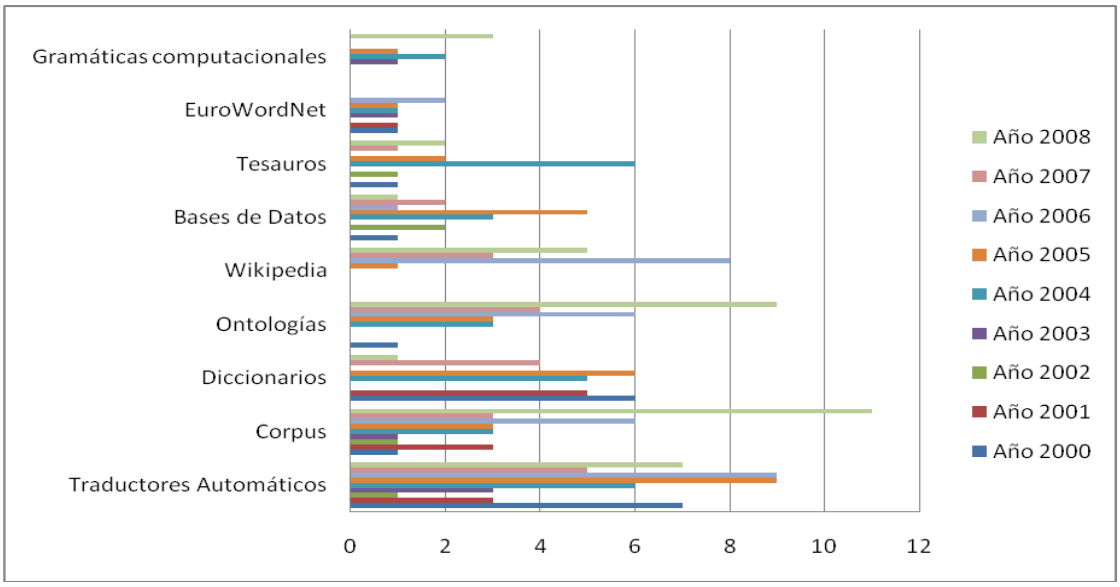

Figura 10. Análisis de todos los recursos y herramientas por años

6. Referencias bibliográficas

Blair-Goldensohn, S. McKeown, K. & Schlaikjer, A.H. (2004). Answering Definitional Questions: A Hybrid Approach. New Directions In Question Answering 4, pp. 47-58.

Crouch, D. Saurí, R. & Fowler, A. (2005). AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines [Online]. Available at: http://www2.parc.com/isl/groups/nltt/papers/aquaint_kb_pilot_evaluation_guide.pdf [Accessed 10 January 2010].

Cui, H. Kan, M.Y. Chua, T.S. & Xiao, J. (2004). A Comparative Study on Sentence Retrieval for Definitional Question Answering. SIGIR Workshop on Information retrieval for Question Answering (IR4QA), Sheffield.

Diekema, A.R. (2003). *Translation Events in Cross-Language Information Retrieval: Lexical ambiguity, lexical holes, vocabulary mismatch, and correct translations.* Tesis Doctoral. University of Syracuse.

Frank, A. Kirefer, H.U. Xu, F. Uszkoreit, H. Crysmann, B. Jörg, B. & Shäfer, U. (2006). Question answering from structured knowledge sources. Journal of Applied Logic, Special Issue on Questions and Answers: Theorical and Applied Perspectives 5, pp. 20-48.

Grefenstette, G. (1998). Cross-Language Information Retrieval. Kluwer academic publishers 1.

Jackson, P. & Schilder, F. (2005). Natural Language Processing: Overview. IN: K. Brown (ed.), Encyclopedia of Language & Linguistics 2, pp. 503-518. Amsterdam: Elsevier Press.

Lee, M. Cimino, J. Zhu, H.R. Sable, C. Shanker, V. Ely, J. & Yu, H. (2006). Beyond Information Retrieval –Medical Question Answering. AMIA. Washington DC, Estados Unidos.

Nguyen, D. Overwijk, A. Hauff, C. Trieschnigg, D.R.B. Hiemstra, D. & de Jong, F.M.G. (2009). WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia. IN: C. Peters et al. (eds.): CLEF 2008, LNCS *5706*, pp. 58-65.

Olvera-Lobo, M. D. & Gutiérrez-Artacho, J. (en prensa y aceptado). Question-Answering Systems as Efficient Sources of Terminological Information: Evaluation. Health Information and Library Journal.

Pérez-Coutiño, M. Solorio, T. Montes y Gómez, M. López López, A. & Villaseñor Pineda, L. (2004). The Use of Lexical Context in Question Answering for Spanish. Workshop of the Cross-Language Evaluation Forum (CLEF 2004).

Tsur, O. (2003). *Definitional Question-Answering Using Trainable Text Classifiers.* Tesis doctoral. University of Amsterdam.

Zweigenbaum, P. (2005). Question answering in biomedicine. IN: De Rijke, M. and B. Webber (eds.), Proceedings Workshop on Natural Language Processing for Question Answering, EACL 2003, 1-4. Budapest: ACL.