EL PAÍS.com Versión para imprimir Imprimir

LINGÜÍSTICA

# La Universidad española, vivero de empresas de las tecnologías de la lengua

## Thera, Eleka y E-genio comercializan aplicaciones desde 2003. Más de 30 grupos universitarios investigan cómo mejorar los sistemas de búsqueda de información en documentos digitales

**LAIA REVENTÓS** 02/09/2004

España cuenta con una "próspera y amplia comunidad científica de las tecnologías de la lengua", aquellas que mediante voz o texto permiten programar ordenadores de manera que se comporten como si entendiesen la lengua humana. Las palabras de Donia Scott, presidenta del Congreso de la Asociación Internacional de Lingüística Computacional (ACL), que celebró el pasado julio su 42º reunión anual en el recinto del Fórum, reflejan el estado de la cuestión en nuestro país de una disciplina científica muy ligada a la inteligencia artificial y que reúne las inquietudes de las humanidades con las de la ciencia y la tecnología.

En España hay más de 30 grupos de investigación, repartidos por las universidades, que trabajan en reconocimiento de voz, procesamiento del lenguaje natural, traducción de texto a texto y síntesis de voz, los cuatro procesos básicos de estas tecnologías. La Sociedad Española para el Procesamiento del Lenguaje Natural, fundada en 1984, agrupa a más de 300, entre socios y empresas. Manuel Palomar, su presidente y catedrático de Lenguajes y Sistemas Informáticos de la Universidad de Alicante, asegura: "Ahora podemos transferir aplicaciones que beneficien a la sociedad. Las de texto son las más avanzadas. Las de voz son más complejas porque es difícil detectar los distintos tonos".

El grupo de Palomar, en colaboración con las universidades de Jaén, la Politécnica de Valencia y la UNED, trabaja para mejorar los actuales sistemas de búsqueda de información, uno de los retos de las tecnologías de la lengua. "Se trata de hacer búsquedas concretas en documentos digitalizados. Es decir que si queremos saber qué mide la torre Eiffel, obtengamos una respuesta concreta ya que la máquina se encarga de la criba de datos". El buscador Tabarca, funciona desde julio con esta tecnología.

### 'Desambiguación'

El problema de estos sistemas "es el tiempo de respuesta. Una de sus soluciones pasa por la *desambiguación* del significado, una técnica que pone cada palabra en su contexto pero que todavía no está resuelta", dice Alfonso Ureña, de la Universidad de Jaén. Su grupo (entre otros) trabaja en sistemas de recuperación de información multilingüe que incorporen técnicas del procesamiento del lenguaje natural; "es decir que incorporan sinónimos, tiempos verbales y realizan análisis sintáctico y semántico tanto en el idioma de la consulta como en otras lenguas".

Otro campo son los sistemas de extracción de información en documentos digitales. "XNotarial es una aplicación que extrae de las escrituras de compraventa los nombres del vendedor, comprador y la finca automáticamente", explica Palomar.

El Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla (TALP), de la UPC, está especializado en el del tratamiento automático del lenguaje natural oral y escrito. "TALP participa con varios grupos europeos en el desarrollo de una agenda electrónica capaz de traducir conversaciones completas del inglés al castellano o catalán", explica Horacio Rodríguez.

Otra consecuencia de la explosión investigadora son las empresas salidas de las universidades. Todas, en 2003. Como Thera, del grupo Clic de la Universidad de Barcelona, que comercializa Ontology, un programa para clasificar documentos de cualquier fuente electrónica a gran velocidad (65.000 palabras por segundo). Además, tiene analizadores morfológicos y sintácticos del catalán, castellano e inglés.

Otro ejemplo es Eleka, surgida del Grupo IXA de la Universidad del País Vasco, que vende un *software* que reconoce textos en euskera mientras están siendo escaneados de un libro. "Eleka utiliza *lematizadores* (programas que detectan la raíz de una palabra); WordNet, diccionario multilingüe donde las palabras están

organizadas por campos semánticos o un corrector ortográfico que se adapta a los procesadores de texto", dice Arantxa Díaz de Ilarraza, de IXA.

E-genio, salida del laboratorio de bases de datos de la Universidad de A Coruña, ha incorporado a la Biblioteca Virtual Galega un sistema que permite buscar cualquier palabra, frase o conjunto de caracteres en cualquiera de las obras almacenadas allí. Nieves R. Brisaboa, directora del laboratorio coruñés, explica que "E-genio también ha digitalizado todos los fondos documentales de la Real Academia Gallega. En breve, su página incorporará la hemeroteca virtual resultante".

SEPLN: www.sepln.org

UA: http://gplsi.dlsi.ua.es

TALP: http://www.talp.upc.es/

THERA: www.thera-clic.com

ELEKA: www.eleka.es

BVG: http://bvg.udc.es/index.jsp

© **Diario EL PAÍS S.L.** - Miguel Yuste 40 - 28037 Madrid [España] - Tel. 91 337 8200
© **Prisacom S.A.** - Ribera del Sena, S/N - Edificio APOT - Madrid [España] - Tel. 91 353 7900