RECURSOS LINGÜÍSTICOS PARA LA COMUNIDAD HISPANOHABLANTE

El español, referencia cultural y activo económico

La práctica totalidad de los hispanohablantes han escuchado o al menos oído que 500 millones de personas hablan español en el mundo, que es la segunda lengua por número de hablantes, el segundo idioma de comunicación internacional y ocupa el segundo puesto entre los idiomas más estudiados en el mundo. La Unión Europea cuenta con cerca de dos millones de estudiantes de educación secundaria, y es el idioma extranjero más popular en los Estados Unidos (elegido por el 70% de los alumnos). También, que ocupa el tercer lugar en número de usuarios en Internet con 175 millones (8%), tras el inglés (580 millones, 28%) y el chino (520 millones, 25%); que dentro de tres o cuatro generaciones el 10% de la población mundial se entenderá en español, y en 2050 Estados Unidos de Norte América será el primer país hispanohablante del mundo.

Sin embargo, la práctica totalidad de los hispanohablantes ni ha escuchado ni, por supuesto, planteado, que el desarrollo tecnológico necesario para apoyar este despliegue es claramente insuficiente, no contribuye a la dinamización del proceso y, en último término, su falta puede poner en riesgo el objetivo final.

«En España existe una pequeña industria lingüística y un marco de investigación que se benefició en el pasado de programas de investigación importantes. (...) Sin embargo, tanto el tamaño de los recursos como el número de herramientas son todavía muy limitados en comparación con los recursos y las herramientas existentes para el inglés, y, desde luego, no son lo suficientemente completos como para dar el apoyo tecnológico integral que necesita una sociedad del conocimiento verdaderamente multilingüe. Desgraciadamente, la implicación de la industria en las tecnologías lingüísticas para el español en la actualidad es reducida. La mayoría de grandes empresas han interrumpido o reducido mucho sus actividades en este área, dejándola mayoritariamente en manos un pequeño grupo de empresas medianas y pequeñas, más especializadas, que no pueden afrontar un mercado internacional en el que la barrera del idioma es un factor clave que frena el comercio electrónico transfronterizo en la UE. Está claro que debe hacerse un mayor esfuerzo para crear recursos lingüísticos para el español, así como impulsar la investigación, la innovación y el desarrollo. La necesidad de grandes cantidades de datos y la gran complejidad de las aplicaciones tecnológicas lingüísticas hace que sea vital el desarrollo de una nueva infraestructura para estimular un mayor intercambio y cooperación» [M. Melero, T. Badía & A. Moreno, Libro Blanco: *La lengua española en la era digital*, MetaNet, Springer, 2012; http://www.meta-net.eu/white papers/e-book/spanish.pdf]. Además de sustento de una cultura prestigiosa y pilar fundamental de la *marca España*, el español debería ser un activo económico de capital importancia al que se prestara máxima atención.

Una nueva infraestructura para la lengua y su industria

Big Data es, hoy en día, ampliamente reconocido como el nuevo paradigma tecnocientífico llamado a tener una extraordinaria importancia en el desarrollo económico de los años venideros. La lingüística y lexicografía, que ya experimentaron en las pasadas décadas un inmenso salto cualitativo gracias a la informática, sin duda van a aprovechar con intensidad esta tecnología emergente. Falta crear, sobre estos nuevos conceptos, una infraestructura ─Norma Digital de la Lengua Española (NDLE)─ que sirva de base para: **A.** Impulsar la investigación lingüística. **B.** Crear nuevas herramientas y recursos para el idioma: algoritmos inéditos de agrupación morfológica, análisis de palabras y análisis distribucional inéditos de propósito general que constituyan tecnología básica que puede ser utilizada en otras aplicaciones, como la agrupación en paradigmas de las palabras desconocidas del corpus. Un novedoso analizador morfosintáctico de textos antiguos que se podrá convertir en herramienta útil para la lematización automática de corpus

diacrónicos con fines didácticos y, especialmente, de investigación. Un sistema de análisis de dependencias para profundizar en el análisis textual utilizando nuevos recursos de análisis, que alcanzará cotas de precisión superiores a lo que existe actualmente. Un novedoso sistema automatizado de categorización y clasificación documental que permitirá enriquecer los textos con atributos que mejorarán sus posibilidades de procesamiento, recuperación y explotación. A diferencia de las clasificaciones manuales, la técnica propuesta producirá clasificaciones consistentes (sin disparidad de criterios ni asunciones apriorísticas). Sistemas automatizados de desambiguación por reglas o estadísticos que permitirán aumentar la calidad de los textos de la NDLE y de cualesquiera otros textos que sean procesados por sus recursos. Una aplicación de comparación y análisis entre diferentes desambiguaciones de un mismo texto (control de versiones). Una aplicación de alineamiento entre el texto desambiguado y las sucesivas anotaciones lingüísticas de un mismo texto por mejora de los recursos y procedimientos de anotación. Un procedimiento innovador para la interconexión relacional y sistemática de datos internos y externos, dado la gran variedad y amplia disponibilidad de fuentes y formatos originales. Y otras muchas aplicaciones relacionadas con el procesamiento de texto y habla. **C.** Desarrollar las industrias de la lengua: movilidad, docencia, tecnologías de texto y del habla. **D.** Blindar la hasta ahora indiscutible preeminencia de España y su Real Academia sobre el idioma español, ante el posicionamiento de otros países (singularmente, México y Estados Unidos) que, apoyados en la universalización de la tecnología y la globalización de los recursos, pueden tener la tentación de disputar esta posición de referencia (Ver: Jean-Baptiste Michel *et al*., «Quantitative analysis of culture using millions of digitized books», *Science* 2011; 331 (6014): 176-82 & Supporting online material).

En resumen, construir la **Norma Digital de la Lengua Española**. La NDLE ha de constar de los siguientes módulos o capas: **A.** ***Banco de datos unificado (BDU),*** que contenga las bases de datos, repositorio de recursos lingüísticos y conectores con otras fuentes lingüísticas. B. ***Capa de tecnologías lingüísticas***, desde donde se realicen procesos de análisis lingüístico y textual que afectarán a las restantes capas (con autonomía respecto a dichas capas) y desde donde se ofrezcan también determinadas funcionalidades de carácter especial a los usuarios de la plataforma. Mayoritariamente, el BDU ha de estar poblado por datos textuales, una de las modalidades de expresión del lenguaje humano. Se pretende asimismo que desarrolle componentes para el tratamiento total de textos: cuestiones formales, procesamiento lingüístico, y análisis de datos. **C**. ***Capa de consulta***, que facilitará el acceso a los usuarios al BDU y a la Capa de tecnologías lingüísticas, permitiendo gran cantidad de criterios de búsqueda (procedencia del texto, naturaleza, autor, etc.) y de navegación sobre el mismo. **D**. ***Capa de servicio***, desde donde se realizarán todas las funciones relacionadas con la gestión y administración de usuarios (identificación, gestión de perfiles, gestión de incidencias, informes de uso, etc.). Soportará además los interfaces (API´s) sobre los que agentes externos podrán desarrollar sus propias aplicaciones, productos y servicios. **E**. ***Entorno de redacción***. Aplicación de redacción específicamente diseñada para la edición de obras lexicográficas y normativas, y que integrará además interfaces de conexión estándares con el mundo exterior. El entorno de redacción debe construirse como una plataforma abierta, modular y adaptable, que soportará todo el ciclo de vida de un proyecto lexicográfico, desde la documentación a la publicación multimodal. Se creará una biblioteca software de componentes reutilizables, basados en estándares y que expongan servicios en la nube. En tanto que componente de la NDLE, el entorno de redacción ha de asegurar la interrelación e interconexión con el resto de componentes, en especial el BDU, de modo que los redactores dispongan de todos los recursos que precisen. **F**. La arquitectura de la NDLE estará soportada horizontalmente por las infraestructuras micro-informáticas necesarias para crear un entorno colaborativo.

Pedro R. García Barreno
Daniel Martín Mayorga
F. Eugenio Martín Fuentes
UIMP - Julio 2013.