JESÚS CARRETERO PÉREZ, CATEDRÁTICO DE ARQUITECTURA Y TECNOLOGÍA DE COMPUTADORES

"Llevo mucho tiempo con la idea de hacer un corrector ortográfico en español de libre distribución"

Jesús Carretero Pérez (Soria, 1963), Catedrático de Arquitectura y Tecnología de Computadores en la Universidad Carlos III de Madrid (UC3M), es coautor del primer corrector electrónico en español de distribución libre (COES). Aunque ingeniero informático de profesión, ha aplicado sus conocimientos técnicos al área de la lingüística. Lo que comenzó siendo un *hobby,* ha acabado siendo el software que los sistemas operativos Linux y Unix incorporan desde 1994 para corregir textos en español. En la actualidad cuenta con unas 70.000 palabras raíz y permanece en constante actualización, alcanzando en 2007 los 182.000 visitantes.

UC3M 2/4/2008 12:43 CEST

Jesús Carretero

Fuente:OIC/UC3M

¿Cómo funciona el COES y qué mejoras presenta frente a otros sistemas de corrección?

COES es una herramienta para corrección ortográfica de textos electrónicos. No es un diccionario de definiciones, como la versión electrónica del diccionario de la Real Academia Española (RAE), sino un generador gramatical que cuenta con una colección de palabras raíz, unas 70.000 aproximadamente, a las que se las aplica una determinada serie de reglas para generar sus palabras derivadas. En la actualidad, contamos con unas 600.000 palabras derivadas, una de las bases de datos de palabras (*corpus*) más grandes que hay en español. También incluye un diccionario de sinónimos, corrección de textos en línea y una búsqueda por contexto.

¿Cómo se le ocurrió crear este corrector?

Todo este tema surgió cuando escribía la tesis doctoral. Utilizaba una máquina en Unix y no había ningún corrector ortográfico para el español. A Santiago Rodríguez, profesor de la Universidad Politécnica de Madrid, y a mí se nos ocurrió la idea de hacer un corrector e integrarlo en las máquinas con Unix. Funcionó y actualmente es el corrector para español integrado en todos los sistemas Unix y Linux del mundo como software de libre distribución. Nunca nos planteamos comercializarlo porque como parte de la Universidad nos apetecía potenciar el software libre y porque pensamos que es importante promover el uso del español en informática, sobre todo con la expansión de Internet.

¿Qué proceso siguieron para seleccionar las palabras raíz y a partir de ahí, alimentar el diccionario?

Al principio, cuando empezamos a trabajar en este tema en 1994, realizamos una lista de palabras raíz a partir del diccionario de la RAE. Para determinar cuáles de ellas eran las más habituales, realizamos búsquedas en textos electrónicos y obtuvimos estadísticas de las que se utilizaban más frecuentemente. Por ejemplo, cuando empezamos a probarlo, usamos palabras antiguas de ediciones electrónicas de El Quijote, libros científicos, es decir, palabras especializadas o antiguas pero correctas y que en muchos casos se siguen utilizando en determinados contextos. Como nuestra capacidad de trabajo era limitada, fuimos añadiendo las más comunes y

lógicamente, las que faltan por añadir en la actualidad son las menos usadas. Con la última versión, la 1.9, el 90% de los usuarios puede corregir sus textos perfectamente teniendo en cuenta que el diccionario de la RAE tenía, en el momento en que empezamos a trabajar en este tema, unas 85.000 palabras raíz y ya hemos incluido unas 70.000 en nuestro corrector. Intentamos seguir ampliándolo para que se mantenga como algo vivo.

¿Y para generar las palabras derivadas?

Para ello, utilizamos una herramienta de libre distribución que está incorporada en los sistemas Linux y Unix llamada Ispell. Siguiendo las reglas del Ispell aplicamos un formato estructurado de generar palabras partiendo de las palabras raíz y luego las derivamos aplicando esta herramienta. Por ejemplo, a partir del verbo cantar, las reglas de derivación para los verbos regulares producen todas sus formas conjugadas, cantaré, cantare, cantaba, cantando, etcétera. Esto implica muchísimo trabajo, y al principio más todavía. Era tan ímprobo, que no encontramos a nadie que nos ayudara. Sacar adelante el proyecto es una cabezonería de nosotros. La gente piensa que el castellano es un idioma muy regular, pero no es verdad, presenta muchísimas irregularidades. En total, trabajamos con unas tres mil reglas de derivación para responder a todas las variantes de las palabras raíz.

¿Cómo puede utilizarse?

El corrector puede usarse de dos maneras. La primera es vía web a través de nuestra página http://www.datsi.fi.upm.es/~coes/. En ella el usuario puede corregir una palabra o bien adjuntar un texto completo de forma totalmente gratuita. No obstante, esta herramienta está incluida en todas las versiones de Linux y Unix del mundo y todas utilizan este corrector de español. De hecho, figuramos en el copyright. El corrector también está preparado para ser incorporado en el sistema operativo Windows (Word), aunque este sistema incluye su propio corrector.

¿Cuántos usuarios tiene COES?

En la actualidad puede tener millones de usuarios. En la página web contamos unos 182.000 visitantes de la página, lo que supone unos 20.000 al año. A esta cantidad hay que sumarle todos los usuarios de Linux y Unix

en el que el COES está incorporado. En la actualidad, curiosamente, estamos recibiendo algunos contactos de empresas interesadas como, por ejemplo, empresas de videojuegos o de móviles. Además, los usuarios, empresas o particulares, tienen la ventaja de que se lo damos gratis.

¿Qué áreas científicas, informáticas, han tenido que desarrollar para hacer el corrector? Supongo que también habrán tenido que repasar cuestiones de gramática española.

Fundamentalmente hemos trabajado con técnicas de compilación de programas, herramientas de programación que entendieran las reglas de derivación que aplicamos con Ispell. Casi todo lo que necesitamos usar de Informática para desarrollar el corrector lo aprendí cuando estudié la carrera, pero gramática sí que he aprendido mucho ya que hemos usado como base el libro de gramática de Lázaro Carreter y la Gramática Española publicada por la RAE. De hecho, estoy muy orgulloso porque mi compañero Santiago y yo tenemos una publicación en 1996 en la Revista Española de lingüística, principal órgano de expresión de la Sociedad Española de Lingüística.

¿Reconoce también palabras en otros idiomas?

El COES reconoce anglicismos o palabras en otros idiomas, como nombres propios, de ciudades, etc. La cuestión es que no puede generar, con las reglas gramaticales que usamos, palabras que no sean españolas, pero sí que las incluimos, y el diccionario las reconoce y son distribuidas en las nuevas versiones. Intentamos sacar actualizaciones cada año, pero en realidad, hasta que no tengamos material que implique un cambio sustancial no sacamos una versión nueva.

¿Cuánto llevan trabajando en el corrector?

Estuvimos trabajando en este tema intensamente cuatro años, entre 1994 y 1998. En la actualidad, nuestra función básica es mantener el corrector actualizado, ampliando la lista de palabras. En ocasiones, los usuarios nos informan de que existe algún error, y también ocurre que algunas palabras caen en desuso o que se reducen a contextos especializados. Estas últimas las tratamos aparte de manera que, tenemos el diccionario castellano general, con unas 70.000 palabras, y también diccionarios temáticos con

términos específicos. Por ejemplo, tenemos uno de medicina que nos gustaría mucho ampliar aunque también sería muy interesante abarcar más temáticas como informática o biología. Como siempre digo, estamos abiertos a todo tipo de colaboraciones.

Los programas de software de distribución libre suelen contar en muchas ocasiones con la participación de los propios usuarios. ¿Cuentan con internautas que colaboren suministrando nuevas palabras?

Sí, tenemos gente que colabora pero como es software libre colaboran, nunca mejor dicho, de forma libre, esporádica. Cualquier usuario en su instalación puede añadir lo que quiera, de hecho nos consta que lo hacen, aunque no nos lo dicen. También hay empresas que lo utilizan y añaden sus propios diccionarios específicos. Pero en el diccionario que se distribuye con Linux y Unix, nosotros tenemos el control centralizado principalmente para que se mantenga el rigor.

¿Cuáles son sus líneas de investigación actuales?

Ahora mismo estamos mejorando el diccionario poco a poco. Aún así, yo llevo mucho tiempo con la idea de hacer un traductor de libre distribución, en primer lugar en inglés, pero por falta de mano de obra lo he dejado aparcado. Hay muchos en el mercado pero creo que se puede aportar algo y sigo con ganas de hacerlo.

¿Está entonces su trabajo orientado fundamentalmente a la lingüística?

En realidad, el tema de la lingüística lo hago por *hobby*, pero no trabajo en eso. Mis líneas de investigación son sistemas informáticos y computadores, y los temas de supercomputación. También arquitectura de computadores, optimización y programación para procesadores multicore, transformación automática... La razón de trabajar en temas de lingüística es que estos temas nos gustan y nos parecen útiles.

Más información:
Oficina de Información Científica de la UC3M

Imagen en alta resolución