Una Herramienta de Recuperación Morfoléxica Aplicada a *Microsoft Word*

Octavio Santana Suárez *(osantana@dis.ulpgc.es)*
Universidad de Las Palmas de Gran Canaria
Zenón Hernández Figueroa *(zhernandez@dis.ulpgc.es)*
Universidad de Las Palmas de Gran Canaria
Gustavo Rodríguez Rodríguez *(grodriguez@dis.ulpgc.es)*
Universidad de Las Palmas de Gran Canaria
Luis Losada García *(llosada@dis.ulpgc.es)*
Universidad de Las Palmas de Gran Canaria

1. Introducción

Uno de los aspectos de la investigación en lingüística es el estudio del uso de la lengua en documentos escritos; se trata de identificar y analizar la aparición de determinadas construcciones, lo que, en gran medida, puede entenderse como una clase particular de lo que en informática se conoce como recuperación de información. En el ámbito de la recuperación de información se ha tenido desde siempre conciencia de la insuficiencia de las búsquedas exacta y parcial de las palabras de un texto, y también de la necesidad de incorporar información lingüística para una recuperación más completa. Las ya antiguas búsquedas con truncamiento parten de la hipótesis de que las distintas formas de una palabra se componen de una raíz fija acompañada de un sufijo o un prefijo variables; tal hipótesis suele ser bastante acertada para lenguas poco flexivas, pero resulta muy pobre con lenguas muy flexivas y con altas tasas de irregularidad. Las búsquedas con máscara, por similitud o en base a expresiones regulares no incorporan la adecuada información sobre la naturaleza morfológica de las palabras.

2. Antecedentes

El *Grupo de Estructuras de Datos y Lingüística Computacional* (*GEDLC*, <http://www.gedlc.ulpgc.es>) del *Departamento de Informática y Sistemas de la Universidad de las Palmas de Gran Canaria lleva* algún tiempo desarrollando trabajos en morfología computacional, sintaxis automatizada, análisis de textos y lexicografía que incluyen lematizadores y flexionadores del español, así como el estudio de relaciones morfoléxicas entre las palabras.

El bagaje de conocimientos acumulado y la experiencia en el desarrollo de herramientas en el campo se ponen en este trabajo al servicio del desarrollo de sistemas de localización de fenómenos morfológicos del español dentro de un texto. Se ha realizado una aplicación de búsqueda lingüística aplicada a un procesador de textos popular —*Microsoft Word*.

El hecho de que el diálogo "Buscar y reemplazar" de *MS-Word XP* ofrezca una opción llamada «Todas las formas de la palabra» que según la ayuda de la aplicación sirve para «Buscar o reemplazar sustantivos, adjetivos o tiempos verbales» demuestra el interés de este tipo de búsquedas en el contexto de un procesador de textos. Pero la propia ayuda de la aplicación hace dudar del alcance de tales búsquedas al poner ejemplos como: «reemplace 'manzana' por 'naranja' y, al mismo tiempo, reemplazará 'manzanas' por 'naranjas'» o «reemplace 'peor' por 'mejor' y, al mismo tiempo, reemplazará 'el peor' por 'el mejor'» ; ambos casos corresponden a simples sustituciones de cadenas de caracteres que no requieren ningún conocimiento lingüístico especial y que, de hecho, se pueden realizar sin seleccionar la opción «Todas las formas de la palabra» ; más prometedor parece el ejemplo de los verbos: «reemplace 'dormir' por 'salir' y, al mis mismo tiempo, reemplazará 'dormido' por 'salido'» , pero el *GEDLC* no ha logrado verlo funcionar.

3. La herramienta desarrollada

Se ha desarrollado una herramienta de búsqueda textual para *MS-Word* que incorpora los aspectos flexivos, derivativos y prefijales entre otros mecanismos de formación de palabras del español, lo que aporta una gran potencia de búsqueda.

A la hora de diseñar una aplicación que permita especificar patrones de búsqueda que tengan en cuenta aspectos flexivos y derivativos de la lengua hay que observar una cuestión fundamental: la gran cantidad de detalles que son susceptibles de configuración —la flexión verbal admite 116 configuraciones diferentes.

3.1 Organización

La aplicación se ha diseñado para presentar distintos niveles de detalle. El nivel básico muestra: una caja de entrada de texto, en la que el usuario debe introducir la palabra a buscar, un botón para iniciar la búsqueda, otro para usar la palabra como parte de una coocurrencia, y un par de flechas que dan acceso a mayores detalles.

Figure 1

El usuario sólo tiene que escribir una palabra y pulsar el botón Buscar. El patrón de búsqueda que se aplicará será el que esté configurado —por defecto corresponde a "cualquier palabra del texto que tenga una forma canónica que coincida con alguna de las formas canónicas de la palabra de búsqueda y que, para esa forma canónica, tenga la misma flexión".

En el siguiente nivel de detalle se pueden elegir los grados de derivación y de flexión; se usan tres escalas independientes: una para la derivación y otras dos para la flexión de las formas verbales y de las no verbales.

Si el usuario requiere una recuperación más precisa accederá al último nivel de detalle de la flexión —las relaciones morfoléxicas continúan en un nivel paralelo—, donde se podrá modificar el patrón de búsqueda, ampliando o recortando elecciones de flexión. Existe la posibilidad de añadir o quitar del patrón de búsqueda prefijos y, en el caso de los verbos, pronombres enclíticos.

Cabe tener en cuenta las formas canónicas que correspondan a la palabra de búsqueda o ignorarlas: por ejemplo, buscar palabras que sean "primera persona del singular del presente de indicativo de un verbo introducido" o, ignorando la forma canónica, "primera persona del singular del presente de indicativo de cualquier verbo".

Eligiendo una forma canónica se accede a la interfaz de configuración de las relaciones morfoléxicas en donde se puede indicar qué formas relacionadas se desea incluir en la búsqueda.

Si el usuario escribe un asterisco en lugar de una palabra, se abre la posibilidad de configurar un patrón de búsqueda por características morfogramaticales, sin determinación léxica; por ejemplo, localizar todas las palabras que sean "sustantivos femeninos plurales" o "formas verbales del presente de indicativo", independientemente de cualquier forma canónica.

Además de la búsqueda de palabras individuales, es posible la localización de coocurrencias, tanto con determinación léxica, como por características morfogramaticales —lo que permite afinar la búsqueda hasta el punto de poder situar fenómenos lingüísticos específicos, tales como: perífrasis verbales, regímenes preposicionales y colocaciones léxicas.

4. Conclusiones

Se ha elegido *MS-Word* por ser, seguramente, el procesador de textos más extendido bajo el entorno *MS-Windows* y disponer de interfaz COM (Component Object Model) que facilita la interoperabilidad con otras aplicaciones. La concepción de la herramienta en si es tal que podría interactuar con otras aplicaciones que ofrezcan interfaces COM.

De hecho, el objetivo principal consistió en cómo configurar una interfaz que aprovechara lo motores de lematización desarrollados por el *GEDLC* para realizar búsquedas que incorporen conocimiento lingüístico de forma potente, usable y efectiva. La decisión de que la herramienta desarrollada se aplicase a un procesador de textos pretendió evitar las distracciones derivadas de problemas particulares de otros ámbitos, tales como los de la navegación, si la herramienta se aplicaba a realizar búsquedas en la red, por ejemplo. El siguiente paso será adaptar la interfaz desarrollada para aplicarla a entornos más complejos que un procesador de textos, tales como: el análisis de corpus, el estudio del uso de la lengua en Internet, herramientas de apoyo a la enseñanza, etc. Es un proceso abordable dada la experiencia que también posee el *GEDLC* en ese campo, como se refleja en trabajos previamente publicados sobre analizadores de páginas Web. Análogamente, la herramienta desarrollada podría aplicarse a la recuperación de información en bases de datos textuales.

Bibliografía

Figuerola, Carlos G., Raquel Gómez, Angel F. Zazo Rodríguez, and José Luis Alonso Berrocal. "Stemming in Spanish: A first approach to its impact on information retrieval." *Results of the Cross-Language System Evaluation Campaign CLEF 2001, Darmstadt, Germany.* Ed. Carol Peters. September 2001. 197-202.

Figuerola, Carlos G., Raquel Gómez Diaz, Angel F. Zazo Rodríguez, and José Luis Alonso Berrocal. "Spanish monolingual track: the impact of stemming on retrieval." *Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Darmstadt, Germany, September 2001; Revised papers, volume LNCS 2406 of Lecture Notes in Computer Science* (2002): 253-261.

Santana, O., F. Carreras, J. Pérez, and G. Rodríguez. "Relaciones morfoléxicas prefijales del español." *Procesamiento de Lenguaje Natural* 32 (2004): 9-36.

Santana, O., F. Carreras, J. Pérez, and G. Rodríguez. "Relaciones morfoléxicas sufijales en español." *Procesamiento de Lenguaje Natural* 30 (Marzo, 2003): 1-73.

Santana, O., J. Pérez, Z. Hernández, F. Carreras, and G. Rodríguez. "FLAVER: Flexionador y lematizador automático de formas verbales." *Lingüística Española Actual* XIX.2 (1997): 229-282.

Santana, O., J. Pérez, F. Carreras, Z. Hernández, J. Duque, and G. Rodríguez. "FLANOM: Flexionador y lematizador automático de formas nominales." *Lingüística Española Actual* XXI.2 (1999): 253-297.