Procesamiento del Lenguaje Natural, núm. 28 (2002), pp. 105-106

Sistema Computacional de Gestión Morfológica del Español (SCOGEME)

Francisco Javier Carreras Riudavets
Universidad de Las Palmas de Gran Canaria
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
fcarreras@dis.ulpgc.es

Director: Octavio Santana Suárez
Universidad de Las Palmas de Gran Canaria
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
osantana@dis.ulpgc.es

Codirector: José Rafael Pérez Aguiar
Universidad de Las Palmas de Gran Canaria
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
jperez@dis.ulpgc.es

Resumen: Esta tesis presenta una aplicación que permite probar la potencialidad de un motor capaz de gestionar la morfología del español. Se tratan los aspectos lexicográficos relacionados con la morfología flexiva y derivativa y la prefijación. Como principal objetivo se establecen relaciones morfoléxicas —90.000 aprox.— entre los vocablos de un corpus de 134.000 formas canónicas o entradas recogidas de los principales repertorios lexicográficos del español. Además, se consideran otras relaciones morfoléxicas entre palabras, no catalogadas dentro de la derivación, muy útiles para el procesamiento del lenguaje natural por poseer características funcionales y semánticas similares a las que se desprenden de los procesos de formación de palabras.
Palabras clave: morfología, lematización, flexión, derivación, prefijación, lingüística computacional.

Abstract: This thesis presents an application that allows to prove the potentiality of a motor able to manage the morphology of the Spanish. It treats the lexicographical aspects related to the flexive and derivative morphology and the prefixation. As a main objective, morpholexical relations —90.000 approx.— between the words of corpus of 134.000 canonical forms gathered of the main lexicographical repertoires of the Spanish are settled down. In addition, we will consider other very useful for the processing of the natural language when having functional and semantic characteristics similar to which they are come off the processes of formation of words morpholexical relations between words, not catalogued within the derivation.
Keywords: morphology, tagged, flexion, derivation, prefixation, computational linguistic.

En esta tesis se desarrolla un sistema capaz responder a cualquier aspecto morfológico de una palabra del español que abarca todo lo relacionado con la morfología flexiva, derivativa y la prefijación, para el establecimiento de relaciones morfoléxicas. Permite el reconocimiento, la generación y la manipulación de las relaciones morfoléxicas a partir de cualquier vocablo, así como el del campo morfoléxico al que pertenece, categoría gramatical de la base y de sus palabras relacionadas, incluye la recuperación de toda su información lexicogenética hasta llegar a una primitiva, la gestión y control de los afijos en el tratamiento de sus relaciones, así como la regularidad en la relación establecida y otros aspectos.

Sobre un corpus bastante amplio —134.109 formas canónicas o entradas—, se organiza un estudio taxonómico, exhaustivo y sistemático de los afijos utilizados en las relaciones morfoléxicas que proporciona una visión global

ISSN 1135-5948

del comportamiento y productividad de las palabras del español en los principales procesos de formación —sufijación, prefijación, parasíntesis, supresión, regresión, modificación-cero, apócope, metátesis y otros no clasificables que generan grafías alternativas— y se establece una descripción pormenorizada de las relaciones entre palabra y afijo en el marco de la funcionalidad, formalidad y semántica.

Se recopilan, por tanto, todos los procesos formativo-derivativos del español, para establecer relaciones morfoléxicas, sin entrar en discusiones teóricas de uso, formalismos históricos y otros aspectos poco prácticos desde el punto de vista de un procesamiento del lenguaje natural dirigido a la automatización de mecanismos lingüísticos, ya sean hablados o escritos. Se pretende establecer sistemas automáticos con una amplia y sencilla capacidad de adaptación a la evolución de la lengua, donde el coste de recuperación ante posibles errores sea mínimo.

Estas relaciones se extienden a vocablos que poseen una relación funcional y semántica similar a la que se establece entre una primitiva y una derivada, y que no necesariamente se ha formado uno del otro. El objetivo de tal ampliación es completar un conjunto de relaciones morfoléxicas que desde el punto de vista del procesamiento del lenguaje natural son necesarias para abarcar todas las posibilidades que ofrece la lengua en este sentido. Al conjunto de todas las relaciones que se persiguen se le denomina *relaciones morfoléxicas extendidas*. Se recopilan 90.000 relaciones morfoléxicas extendidas que dan lugar a unas 30.000 familias de palabras que conforman 14.000 clanes distintos.

Se cataloga ampliamente el conjunto de palabras del español que pertenecen al mismo campo morfológico —idéntico afijo y categorización gramatical. Se dispone por tanto de una visión total de la productividad y del comportamiento de un afijo en la morfología derivativa y prefijal, así como su regularidad.

Tanto desde el punto de vista funcional de la palabra original como del de su relacionada, se realiza un estudio completo de la morfología derivativa y compositiva —limitada a la prefijación— para todas las categorías gramaticales: verbos, sustantivos, adjetivos, adverbios, pronombres, preposiciones, conjunciones, expresiones y otras. Se conoce la distribución transcategorizacional, la productividad y el comportamiento del uso de las categorías gramaticales en la morfología derivativa y prefijal.

Las relaciones morfoléxicas extendidas y los resultados obtenidos, junto con la herramienta FLAPE —Flexionador y Lematizador Automático de Palabras de Español—, se integran para dar lugar a un Sistema Computacional de Gestión Morfológica del Español (SCOGEME) capaz de gestionar simultáneamente la morfología flexiva y la derivativa, los prefijos y la parasíntesis, y las relaciones entre palabras del español con un punto de encuentro en su historia etimológica.

Se obtiene una aplicación diseñada para ser de utilidad a quienes tratan con documentos en español: lexicólogos, analistas de estilo, recuperadores de información textual, traductores, etc. Una intuitiva interfaz gráfica, con manejo de ventanas de diálogo, botones y demás herramientas, facilita la interacción persona-ordenador. Esto supone un primer paso hacia las múltiples posibilidades informáticas y aplicaciones especializadas que deben desarrollarse sobre esta base de conocimiento.

SCOGEME resulta un producto fácilmente integrable en herramientas de ayuda al tratamiento de documentos orientadas a resolver problemas del procesamiento del lenguaje natural —corrector ortográfico, buscador avanzado de información, analizador de textos, desambiguador, estación lexicológica, analizador sintáctico, extractor de información, generador automático de textos, corrector sintáctico, extractor de resúmenes, etc.

Este motor morfológico se ha implementado en lenguaje C++ y está disponible para plataforma Windows y Linux. Sobre el sistema operativo Windows se ha diseñado una aplicación de usuario que permite observar la potencialidad y las posibilidades de dicho trabajo, y sobre Linux se ha implementado una interfaz Web, con un menú de opciones simplificado respecto de la aplicación Windows, que hace posible su acceso a través de Internet de forma discrecional y gratuita en la página http://gedlc.ulpgc.es. Se acompaña de una presentación gráfica de las relaciones morfoléxicas extendidas solicitadas.