Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 369-370 recibido 1-05-2009; aceptado 5-06-2009

OAL: una herramienta para el desarrollo de recursos lingüísticos

OAL: a tool for the development of linguistic resources

Javier Couto[1,2], Helena Blancafort[1,3], Somara Seng[1], Anass Talby[1], Claude de Loupy[1,2]

[1] Syllabs
Paris, France
{couto,blancafort,loupy}@syllabs.com

[2] Laboratoire Modyco
Université de Paris 10
Nanterre, France

[3] Universitat Pompeu Fabra
Barcelona, España

Resumen: OAL es un conjunto de herramientas para ayudar a los lingüistas a crear, alimentar y gestionar léxicos computacionales. Está diseñado como un cliente-servidor que incluye un agente inteligente de aspiración de contenido textual Web y una cadena para trabajar a partir de corpus bruto.
Palabras clave: editor de léxicos, aspirador de contenido textual Web.

Abstract: OAL is a tool to assist the linguist in the development, enrichment and management of computational lexicons. It has been designed as a client-server which includes an intelligent web crawler and a chain to process raw corpora.
Keywords: lexicon editor, web crawler.

1 Introducción

Uno de los aspectos claves en el procesamiento de lenguaje natural está vinculado con la calidad de los recursos lingüísticos (p.ej. léxicos morfosintácticos) sobre los cuales se basan las aplicaciones. Aún hoy, su creación y enriquecimiento suponen un proceso costoso y, muchas veces, tedioso.

Sin herramientas específicas, quien desarrolla estos recursos necesita mancomunar los resultados de distintos tratamientos: aspiradores de contenido Web, *scripts* para lidiar con problemas de codificado e idioma, *scripts* para extraer el contenido textual, para realizar importaciones y exportaciones de información, entre otros. La información suele guardarse en formato textual y la manipulación de varios ficheros es una práctica corriente.

Para resolver esta problemática, Syllabs dispone de OAL[1], un conjunto de herramientas integradas para ayudar al lingüista y, especialmente, favorecer el control de calidad mediante procedimientos automáticos para evaluar cada recurso, medir la regresión, verificar cierta armonía y coherencia entre los diferentes idiomas, gestión de las particularidades de los recursos multilingües.

[1] La sigla OAL corresponde a "Outils d'Aide aux Linguistes", es decir "Herramientas de Ayuda a los Lingüistas".

OAL ha sido diseñado como un frame-work cliente-servidor compuesto de:

- un servidor de recursos lingüísticos
- editores para la gestión de léxicos
- un agente inteligente de aspiración de contenido textual del Web
- una cadena de procesamiento para alimentar léxicos a partir de corpus brutos
- un *guesser* de paradigmas de flexión

La demo presenta el editor de léxicos morfosintácticos, sitúa esta herramienta en la arquitectura general de OAL y comenta los beneficios obtenidos por los lingüistas.

2 Editor de léxicos morfosintácticos

2.1 El formalismo SylLex

El formalismo SylLex permite modelar un léxico morfosintáctico. Este es concebido como un conjunto de lemas –simples y compuestos–, un conjunto de paradigmas de flexión y un conjunto de plantillas de paradigmas de flexión denominados *patterns*.

Un paradigma de flexión reagrupa un conjunto de reglas de flexión y genera las formas asociadas de un lema. La aplicación del paradigma "V/ar" al lema *comerciar*, p.ej., (cf. figura 1) genera todas las formas (flexiones) del verbo *comerciar*. Cada forma tiene asociada una etiqueta morfosintáctica que indica su categoría y atributos, como número o modo. Las formas se calculan a partir de fórmulas que

utilizan expresiones regulares y son definidas por el usuario en el paradigma. En el caso del castellano, las fórmulas son generalmente de tipo "*stem* + sufijo". El valor del *stem* es calculado automáticamente a partir del lema. De esta manera, un paradigma de flexión puede aplicarse a todos los lemas pertinentes y las formas serán calculadas automáticamente. Además, SylLex permite definir variantes léxicas de una forma. Pueden ser de distintos tipos (ortográfica, geográfica, etc.). Una forma puede tener más de una variante.

Los paradigmas de flexión se crean a partir de un *pattern*, una plantilla que indica todas las formas que tiene un verbo con sus etiquetas correspondientes. Un paradigma de flexión es una instancia concreta de un *pattern* en la que el usuario indica la regla de flexión para generar cada forma. Además, el formalismo ofrece un mecanismo de herencia de paradigmas para crear paradigmas similares, como el paradigma de verbos defectivos en "ar" para "tronar" a partir del paradigma de verbos en "ar".

Los paradigmas de flexión para lemas compuestos están representados como combinaciones de los paradigmas de flexión de los lemas simples que lo componen. De esta manera el usuario se ve dispensado de la creación de paradigmas compuestos y las repercusiones de los cambios operados sobre los paradigmas simples se pueden realizar automáticamente. El usuario debe indicar el núcleo del compuesto salvo que sea una excepción.

El modelado de SylLex permite representar un léxico morfosintáctico de forma compacta, consistente y lingüísticamente fundamentada.

2.2 Editor de léxicos

La figura 1 presenta la pantalla principal de edición de un léxico morfosintáctico. A la izquierda se ve una lista con los lemas simples y compuestos del léxico cargado. Cuando el usuario se posiciona sobre un lema, el panel de la derecha muestra los paradigmas de flexión aplicados al lema. En la figura, el usuario ha realizado una búsqueda por "comer" –el sistema se posiciona automáticamente y colorea los elementos pertinentes de la lista–, luego ha seleccionado "comerciar". En el panel de la derecha aparece el paradigma de flexión "V/ar", aplicado al verbo comerciar.

El paradigma "V/ar" contiene tres *stems* posibles y una variante morfológica para cada forma del pretérito imperfecto del subjuntivo. En la parte inferior del panel de edición, se observan las formas y variantes generadas para el lema comerciar. El usuario puede modificar el paradigma, eliminar su aplicación y elegir otro paradigma, crear o suprimir paradigmas. En el caso de los lemas, el usuario puede introducirlos manual- o automáticamente utilizando la función de importación. El sistema ofrece la posibilidad de filtrar las entradas por diversos criterios: lemas, formas, paradigmas aplicados, etiquetas morfosintácticas. Un filtro se visualiza en pantalla y el resultado puede ser exportado a diversos formatos.

Además, OAL incluye un agente inteligente de aspiración de textos Web para encontrar palabras nuevas y así alimentar los léxicos. Un *guesser* de paradigmas propone al lingüista uno o varios paradigmas para cada una de esas palabras a incluir en el léxico.

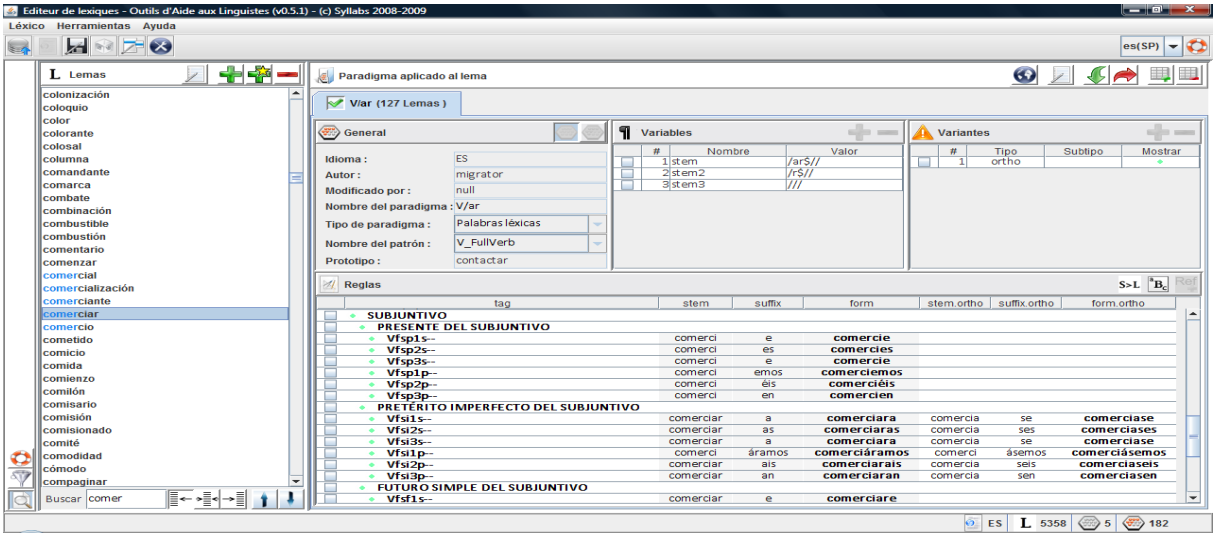

Figura 1: Pantalla principal de edición de un léxico morfosintáctico.