Lluís de Yzaguirre Maura
Anna Matamala Ripoll
M. Teresa Cabré Castellví

Institut de Lingüística Aplicada
Universitat Pompeu Fabra
de_yza@upf.es
www.iula.upf.es

EL LEMATIZADOR "PALIC" DEL IULA (UPF)
(XVII Congreso AESLA, Lingüística del Corpus y Computacional)

ABSTRACT

El PALIC (Programa de Atribución de Lemas y Categorías) es un analizador morfológico flexivo multilingüe que se aplica en la lematización y etiquetaje gramatical de los textos en castellano del corpus técnico del Institut Universitari de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra (UPF), después de las fases de marcaje estructural y preproceso. En esta comunicación damos una visión general del funcionamiento del programa, centrándonos en la generación y mantenimiento de los diccionarios, en la manipulación del texto (o *text-handling*) y en la lematización propiamente dicha. También indicamos otras posibles aplicaciones del PALIC (detección de neología y extracción de terminología) y apuntamos mejoras previstas para el futuro.

1.- Introducción

El PALIC (Programa de Atribución de Lemas y Categorías) es un analizador morfológico flexivo multilingüe que, en el corpus del IULA, se utiliza para lematizar y etiquetar gramaticalmente los textos en castellano. En esta comunicación pretendemos presentar de un modo general este programa: en primer lugar, describiremos su funcionamiento y ámbito de uso; a continuación, explicaremos los tres grandes bloques que lo forman --el preanálisis o generación de diccionarios, la manipulación del texto y la lematización-- y, finalmente, apuntaremos posibles mejoras de cara al futuro.

2.- Descripción y ámbito de uso del PALIC

Como hemos dicho, el PALIC es un analizador morfológico flexivo que se aplica en el tratamiento de los documentos en español del corpus técnico del IULA[1], cuando ya han sido marcados estructuralmente[2] y preprocesados. También ofrece la posibilidad de analizar textos que no han sido preparados para este corpus, lo cual amplia su ámbito de aplicación. Así

[1] Véase Bach, C. *et al.* (1997).
[2] Véase Vivaldi, J. *et al.* (1996).

pues, el analizador recibe un texto tratado o no para el corpus del IULA y genera un documento que contiene información codificada sobre todos los lemas y categorías posibles de cada unidad léxica, de manera que en muchas ocasiones a una misma palabra se le atribuyen varios lemas y categorías. En el PALIC se utilizan unos criterios de lematización establecidos para el corpus del IULA y unos etiquetarios[3] elaborados a partir de las directrices del proyecto EAGLES. Una vez analizados morfológicamente (o sea, lematizados y categorizados) los textos, se les aplica un desambiguador lingüístico AMBILIC y uno de estocástico para resolver las ambigüedades.

El PALIC se empezó a desarrollar en febrero del 1995, pero fue en abril del 1997 cuando se hicieron las primeras pruebas en el corpus del IULA. En diciembre del 1999, la cantidad de palabras procesadas por el PALIC en los documentos en español del corpus técnico del IULA es de casi seis millones. Por áreas temáticas, la distribución es la siguiente:

Derecho	2.048.939
Medio Ambiente	706.651
Informática	665.087
Medicina	1.304.600
Economía	699.776

Sin embargo, el ámbito de aplicación del PALIC no se limita únicamente al análisis morfológico del corpus técnico del IULA. También se ha utilizador para detección de neología[4] y para evaluar estrategias de extracción de terminología[5].

3.- Módulos y fases del PALIC

3.1.- Preanálisis o generación de diccionarios

El PALIC dispone de diccionarios estructurados en módulos. En el caso del castellano, lleva incorporado el lemario del *Diccionario Actual de la Lengua Española* (1995), cedido por Vox-Bibliograf mediante convenio, así como un módulo de neologismos. Para el catalán, dispone de un módulo equivalente a la suma del *Diccionari de la Llengua Catalana* (1983) de Enciclopèdia Catalana y la edición del Fabra de EDHASA de 1980 y varios módulos posteriores con las incorporaciones del *Diccionari de la Llengua Catalana* (1993), del *Diccionari de la Llengua Catalana* (1995) del Institut d'Estudis Catalans y se está preparando el del *Gran Diccionari de la Llengua Catalana* (1998) de Enciclopèdia Catalana. Como vemos, pues, cuando se edita una nueva versión de un diccionario, no se añade al

[3] Véase Morel, J. *et al.* (1997)
[4] Véase Cabré *et al.* (1995)
[5] Véanse Estopà, R. *et al.* (1998a), (1998 b) y (1999).

programa un módulo con todo el diccionario, sino que se modifican los errores de la versión anterior y las novedades se agrupan en un fichero aparte.

En cuanto a otras lenguas, se están elaborando versiones para el francés y el portugués, y también disponemos de listas significativas para el alemán, el italiano, el latín y el holandés (de 20.000 a 50.000 lemas). El hecho de que el PALIC sea multilingüe permite que el usuario pueda buscar unidades en una lengua distinta a la del texto. Por ejemplo, puede ser que en un artículo periodístico escrito en catalán aparezcan voces en castellano: teniendo en cuenta esto, el usuario puede decidir que el programa busque automáticamente las unidades en el diccionario castellano cuando no aparezcan en el catalán, lo caul reduce notablemente el trabajo del neólogo.

Por otro lado, gracias a esta concepción modular, el usuario puede utilizar todos los diccionarios o bien establecer unos filtros y lematizar el texto según una fuente lexicográfica concreta o según los diccionarios de una etapa cronológica determinada, para estudios diacrónicos de neología.

El *mantenimiento de los diccionarios* del PALIC se lleva a cabo en tres fases:

(a) en primer lugar, el administrador introduce palabras en el lemario, con la categoría gramatical correspondiente y un número que codifica el paradigma de la palabra;

(b) cuando hay una cantidad de modificaciones significativa, se ejecuta un programa que genera automáticamente todas las formas de cada lema y las guarda en una base de datos de formas, en la que encontramos varias informaciones codificadas: un vínculo al lema, y a la etiqueta que expresa el vínculo entre la forma y el lema;

(c) una vez generadas todas las formas se acumulan en un mismo registro las idénticas, conservando toda su información.

Respecto a las estrategias clásicas de lematización que analizan en tiempo real, la nuestra requiere una cantidad notable de megaoctetos y mucho acceso a disco, pero se ahorra mucho tiempo de proceso.

3.2.- Manipulación de texto (t*ext-handling)*

El primer bloque del PALIC es la llamada manipulación de texto o *text-handling*, en la que el programa implementa dos comportamientos:

(a) Cuando se aplica al corpus del IULA[6], procesa textos marcados estructuralmente y preprocesados. Por lo tanto, es importante que no altere el marcaje añadido a estos

[6] Para más información sobre el corpus del IULA y las herramientas de tratamiento informático de éste, remitimos a la URL "http://www.iula.upf.es/corpus/corpus.htm".

textos, codificados en SGML y, luego, multiplataforma.

(b) Cuando se aplica a otros tipos de textos, el PALIC es capaz de ofrecer un documento procesable resolviendo cuestiones como el marcaje estructural o el preproceso (fechas, números, locuciones, abreviaturas, etcétera) que en los textos del corpus del IULA ya se han solucionado en fases previas. Sin embargo, al contrario de los textos del corpus, se trata de un marcaje estructural de bajo nivel, ya que no hay tipificación ni jerarquización de párrafos. Los textos que admite pueden estar en ASCII de Macintosh, de MS-DOS (página de códigos 850) y de Windows.

En la fase de *text-handling*, el programa efectúa una lectura de todo el texto, busca la frontera entre las palabras y genera tres listas:

(i) *Lista de formas*: busca las ocurrencias y produce una lista de formas (unidades léxicas y delimitadores) una sola vez.

(ii) *Lista de palabras*: elabora una lista de lo que considera palabras y cada palabra corresponde a un registro que contiene informaciones distintas.
   (a) una cifra que indica el elemento de la lista de formas al que nos referimos;
   (b) el número de lemas que puede tener la palabra y una relación de los lemas con las correspondientes etiquetas para expresar la categoría gramatical..

(iii) *Lista de nodos*: esta lista está formada tanto por elementos léxicos como por delimitadores y se genera para poder reconstruir el texto según el orden original. Dentro de cada nodo (es decir, dentro de cada elemento léxico o delimitador), se incluye la información siguiente:
   (a) una cifra que remite al elemento de la lista de palabras a la que pertenece el nodo (que, a la vez, tiene un número que indica la unidad correspondiente de la lista de formas).
   (b) el *offset*, una variable que permite saber en qué lugar del documento original se encuentra cierta unidad, tomando como punto de referencia el principio del texto;
   (c) las variables que se necesitarán para controlar los pasos de lematización y de desambiguación posteriores y guardar sus resultados..

3. 3.- Lematización

El segundo bloque del PALIC ejecuta la lematización recorriendo el texto para localizar las palabras, buscarlas en los diccionarios preanalizados y, en caso de hallarlas, añadirles lema(s) y etiqueta(s).

Además puede adaptarse a otras situaciones, que detallaremos a continuación:

(i) como hemos comentado en el preanálisis, se pueden seleccionar los parámetros de las

fuentes de validación. Es decir, el usuario puede decidir que quiere lematizar según unos diccionarios concretos o según las obras de un periodo cronológico determinado; un ejemplo de validación peculiar permitió detectar la mayoría de las ambigüedades segmentales del castellano[7].

(ii) se pueden seleccionar otras lenguas para detectar estranjerismos;

(iii) al trabajar con un conglomerado de bases de datos, PALIC puede ser adaptado a procesos peculiares añadiendo campos "ad hoc":

(a) por ejemplo, en un proceso *text-to-speech*, podríamos tener información sobre irregularidades de la pronunciación (tipo "*hegeliano*" o "*freudiano*");

(b) para lematizar textos antiguos, podríamos tener variantes ortográficas;

(c) para recuperación de información, procesamientosemántico o extracción de terminología, podríamos añadir marcas temáticas.

PALIC trata algunos problemas de ambigüedad que interfieren en la lematización:

(i) desambigua las contracciones. En un caso como "*al*", el programa desambigua automáticamente que "*a*" es preposición y "*el*", artículo.

(ii) desambigua los grupos de forma verbal más pronombre enclítico, del tipo "*cásate*".

(iii) marca con información adecuada los casos de ambigüedad segmental como "*pésame*", "*consig*o" o "*verse*", por ejemplo, que serán resueltos por AMBILIC[8].

4.- Posibles mejoras del PALIC

En el futuro está previsto realizar varias mejoras en el programa:

(i) Además de las versiones para Macintosh y MS-DOS ya existentes, se está trabajando en una versión para gpc de Linux, que podría ir seguida de versiones para otros dialectos de Unix.

(ii) El PALIC tendría que ser capaz de utilizar la estructura modular de los diccionarios para encapsular automáticamente con la etiqueta "<foreign>"[9] y la lengua correspondiente las unidades de otras lenguas, sin que fuera necesario marcarlas manualmente.

(iii) Tendrían que incluirse módulos de otras lenguas en el programa.

(iiii) Se tendría que poder operar con documentos más extensos que las muestras del corpus del IULA, que tienen aproximadamente cinco mil ocurrencias.

---

[7] Véase De Yzaguirre *et al.* (2000b)

[8] Véase De Yzaguirre *et al.* (2000a)

[9] Véase Vivaldi, J. *et al.* (1996).

(v) Se tendría que conseguir que los distintos módulos se convirtieran en un conjunto de aplicaciones menores implementadas en Java que funcionaran en procesos cooperativos en red. El objetivo sería que el lematizador y el desambiguador pudieran trabajar en paralelo ya que, de este modo, se aceleraría el proceso.

6.- Bibliografía

BACH, C. ; SAURÍ, R.; VIVALDI, J. y M.T. CABRÉ (1997) "El Corpus de l'IULA: descripció", *Papers de l'IULA*, sèrie informes, 17, Barcelona: IULA, Universitat Pompeu Fabra.

CABRÉ, M. T. y L. DE YZAGUIRRE (1995) "Stratégie pour la détection semiautomatique des néologismes de presse", Traduction, Terminologie, Rédaction, Vol. VIII, n. 2

CABRÉ, M.T. y R. ESTOPÀ (en prensa) Extraction de terminologie: vers un système multifonctionnel, Actas de la I Conferencia sobre la cooperación en materia de terminología en Europa, París.

DE YZAGUIRRE, L., A. MATAMALA, C. BACH, N. CASTILLO y E. USTRELL (2000a): AMBILIC, el desambiguador lingüístico del corpus del IULA (UPF), comunicación presentada al XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

DE YZAGUIRRE, L., S. TORNER Y A. MATAMALA (2000b): El tratamiento automático de las ambigüedades segmentales del castellano, comunicación presentada al XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

ESTOPÀ, R.; VIVALDI, J. y M.T. CABRÉ (1998) "Detectors automàtics de (candidats) a terme: estat de la qüestió", *Papers de l=IULA*, Sèrie Informes, 22, Barcelona: IULA, Universitat Pompeu Fabra.

ESTOPÀ, R. y J. VIVALDI (en prensa) "État de la question des systèmes d'extraction automatique de candidats à terme: vers une proposition intégratrice", Actes de les VII Journées ERLA-GLAT, Université de Brest , p. 385-410.

MOREL, J.; TORNER, S; VIVALDI, J; DE YZAGUIRRE, Ll. y M.T. CABRÉ (1997) "El corpus de l'IULA: etiquetaris", *Papers de l'IULA*, sèrie informes, 18, Barcelona: IULA, Universitat Pompeu Fabra.

VIVALDI, J.; DE YZAGUIRRE, Ll.; SOLÉ, X. y M.T. CABRÉ (1996) "Marcatge estructural i morfosintàctic del Corpus Tècnic amb l'estàndard SGML", *Papers de l'IULA*, sèrie informes, 1, Barcelona: IULA, Universitat Pompeu Fabra.