Procesamiento del Lenguaje Natural, nº39 (2007), pp. 279-280 recibido 07-06-2007; aceptado 22-06-2007

HistoCat y DialCat: extensiones de un analizador morfológico para tratar textos históricos y dialectales del catalán

Jordi Duran Cals
THERA SL
Adolf Florensa s/n
08028-Barcelona
jordi.duran@thera-clic.com

Mª Antònia Martí Antonín
Universitat de Barcelona
Gran Vía 585
08007-Barcelona
amarti@ub.edu

M. Pilar Perea Sabater
Universitat de Barcelona
Gran Vía 585
08007-Barcelona
mpilar.perea@ub.edu

Resumen: Los textos históricos y dialectales del catalán no se pueden anotar morfosintácticamente de manera automática ya que no existe una variante estándar de referencia que permita un tratamiento homogéneo y sistemático. El objetivo de los proyectos HistoCat y DialCat ha sido desarrollar un entorno de anotación semiautomático aprovechando herramientas existentes para la anotación morfosintáctica de textos en catalán, que minimizara al máximo la anotación manual.
Palabras clave: Corpus historicos y dialectales, Anotación Morfosintáctica, Lingüística de Corpus.

Abstract: Catalan historical and dialectal texts cannot be morphosintactically annotated in an automatic way, because there is not a reference standard of written language that could allow a sistematic and homogeneus treatement. The main objective of DialCat and HistoCat projects has been to develop an environment for the semiauthomatic annotation of these corpora using already existing morphological analyzers for standard Catalan trying to minimize the manual annotation.
Keywords: Morphosintactic Annotation, Corpus Linguistics.

1 Introducción. Motivación

Los textos históricos y dialectales del catalán no se pueden anotar morfosintácticamente de manera automática ya que no existe una variante estándar de referencia que permita un tratamiento homogéneo y sistemático.

La anotación morfosintáctica de estos corpus se ha realizado, hasta el momento, de manera manual por no existir un sistema de anotación y lematización automático o semiautomático disponible (Albino, 2006) .

En la lengua antigua, por no existir una variedad estándar de referencia nos encontramos con una gran multiplicidad de formas ortográficas para una misma palabra. En el caso de las variantes dialectales, tenemos que afrontar el problema de determinar como se transcriben ortográficamente las formas propias de ciertas áreas dialectales, que no tienen representación en los diccionarios de la lengua. Es una realidad que la tradición lexicográfica cuenta con muy poca representación dialectal.

El objetivo de los proyectos HistoCat y DialCat ha sido doble. por un lado, se pretendía desarrollar una herramienta para el análisis morfosintáctico automático de textos históricos y dialectales del catalán; por otro, se pretendía recopilar el léxico de la lengua antigua y un léxico dialectal actual, a partir de corpus.

El corpus de la lengua antigua (HistoCat) consta de 97.603 palabras y está formado por textos del siglo XIV, XV y XVI. El corpus dialectal (DialCat) está formado por 23 textos orales en versión fonoortogràfica (cf. Viaplana y Perea, 2003) que presentan variedades locales correspondientes a los seis grandes dialectos del catalán y consta de 36.450 palabras.

Los proyectos que se presentan han consistido en el desarrollo de un entorno de anotación semiautomático aprovechando

ISSN: 1135-5948

herramientas existentes para la anotación morfosintáctica de textos en catalán, que minimizara al máximo la anotación manual.

2 Tratamiento lingüístico de los corpus

Además de la información morfosibtáctica básica que corresponde a la PoS, en el corpus histórico se da información sobre el siglo, la obra y el autor. En los corpus dialectales se indica el dialecto, la variante dialectal y el informante. El anotador puede indicar también si una palabra és un derivado, un péstamo de otra lengua, o bien un barbarismo.

3 Características tecnológicas

El sistema de análisis semiautomático se basa en una versión extendida del analizador HS-Morfo[1]. El sistema de análisis se compone de tres módulos: 1) El etiquetador con el sistema de anotación estándar. 2) El etiquetador con el sistema de anotación histórico/dialectal. 3) La interfaz de validación

3.1 Etiquetador estándar

Este módulo se compone del etiquetador con el sistema de anotación de la lengua estándar, el analizador HS-Morfo. Es el primer módulo en el procesamiento y recibe como entrada el texto plano para crear un documento con el texto segmentado y anotado: cada forma recibe los distintos lemas y etiquetas PoS que puede tener asociados. Aquellas palabras que no reconoce por pertenecer al léxico histórico o dialectal son tratadas en el módulo siguiente.

3.2 Etiquetador con el formario histórico /dialectal

En este segundo módulo se completa la anotación de las formas específicas del vocabulario histórico o dialectal, tanto las formas que no han sido reconocidas en el módulo de análisis estándar, como también aquellas formas que sí se han reconocido pero son ambiguas y pueden recibir nuevas interpretaciones..

[1] HS-Morfo es un analizado cedido por la empresa THERA SL para el desarrollo del proyecto. El desarrollo tecnològico ha sido llevado a cabo por dicha empresa.

3.3 La interfaz de validación

Este último módulo cumple una doble función. Por un lado, el usuario valida qué par lema-PoS de cada forma detectada en los dos módulos previos es la correcta en su contexto. Por otro, permite incluir información nueva, en concreto nuevos pares lema-PoS a aquellas palabras que no se han analizado en los módulos anteriores.

Esta información, una vez introducida pasa a formar parte del sistema de anotación del segundo módulo, el que detecta las formas históricas o dialectales. De esta forma el formario histórico y dialectal se van realimentado de manera que está disponible para futuros tratamientos.

4 Extensiones del sistema

Este sistema es fácilmente extensible a otras lenguas, si se dispone de un analizador morfológico de la lengua estándar.

Actualmente se esta desarrollando una interfaz web de consulta que permitirá recuperar el léxico por los criterios aplicados en el proceso de anotación.

5 Agradecimientos

DialCat (HUM2005-24445-E) e HistoCat (HUM2005-24438-E) son dos proyectos financiados por el Ministerio de Educación en el programa de Acciones Complementarias.

Bibliografía

Albino Pires, Natalia (2006) ‘ULISES: un Integrated Development Environement desarrollado para la anotación de un corpus romancístico’. *Procesamiento del Lenguaje Natural*, n. 37. Septiembre 2006.

Viaplana, J. y Perea, M. P. 2003. *Corpus oral dialectal. Una selecció*. Barcelona. PPU.