Procesamiento del Lenguaje Natural, num. 29 (2002), pp.309-310 recibido 11-07-2002; aceptado 15-07-2002

CATCG: Un sistema de análisis morfosintáctico para el catalán

Àlex Alsina, Toni Badia, Gemma Boleda, Stefan Bott,
Àngel Gil, Martí Quixal, Oriol Valentín
GliCom
Universitat Pompeu Fabra
La Rambla 30-32
08002 Barcelona
{alex.alsina, toni.badia, gemma.boleda, marti.quixal, oriol.valentin}@trad.upf.es,
{stefan.bott, angel.gil}@iula.upf.es

Resumen: CATCG es un sistema de análisis morfosintáctico superficial para el catalán, basado en el formalismo *Constraint Grammar*, que contiene tres herramientas básicas: un analizador morfológico, un etiquetador morfológico y un analizador sintáctico superficial.
Palabras clave: análisis sintáctico superficial, etiquetaje morfológico, catalán

Abstract: CATCG is a shallow parser for Catalan. It uses the *Constraint Grammar* formalism and contains three basic tools: a morphological analyser, a POS tagger and a shallow parser.
Keywords: shallow parsing, POS tagger, Catalan

1 Descripción

CATCG es un sistema de análisis morfosintáctico superficial para texto no restringido en catalán. Es de base lingüística (formalismo *Constraint Grammar*) y altamente modular. Está siendo desarrollado por el grupo GLiCom (Grup de Lingüística Computacional) de la Universitat Pompeu Fabra (Barcelona).

El núcleo del sistema (v. Fig. 1) lo forman tres gramáticas regulares escritas en el formalismo *Constraint Grammar*. El sistema se completa con un módulo de preproceso (verticalización e identificación de oraciones, párrafos, fechas, cifras, nombres propios y abreviaturas) y uno para la proyección morfológica. La proyección no tiene en cuenta el contexto: se proyectan todas las lecturas para cada forma, según la información de un formario (tabla de formas) construido a partir de un analizador-generador morfológico de dos niveles (CATMORF). CATMORF contiene datos sobre categoría morfológica y rasgos flexivos, así como de subcategorización verbal.

1.1 Gramáticas CG

La estrategia esencial de las gramáticas CG consiste en elaborar un análisis morfosintáctico parcial a partir de la información contextual proporcionada en cada oración. Las gramáticas realizan las tareas siguientes:

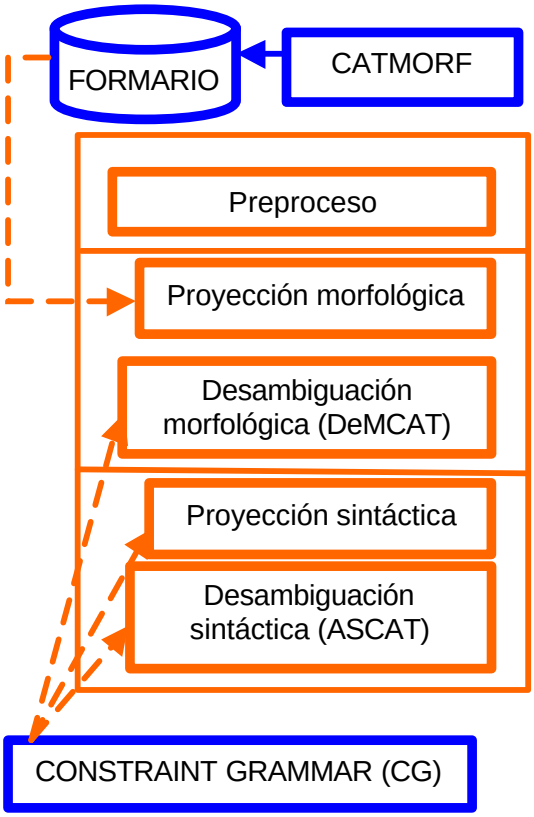

Figura 1: Descripción

A. **Desambiguación morfológica**: esta gramática (DeMCat) elimina las lecturas proyectadas que no se corresponden con el

ISSN 1135-5948

contexto de uso. (1) muestra un ejemplo de regla:

(1) REMOVE TARGET (VERB) IF (0 NOM) (-1 DET) (-2C PREP) ;

B. **Proyección sintáctica**: se realiza de manera controlada, es decir, evitando proyectar lecturas ambiguas en contextos suficientemente seguros.

(2) MAP (@ATR) TARGET (ADJ) IF (-1 VCOP) (NOT *1 NOM BARRIER BAR-DF OR COMA) ;

C. **Análisis sintáctico superficial** (ASCat): proporciona información sobre la función sintáctica de cada palabra (no sobre constituencia): se asigna una etiqueta con el nombre de la función y, en algunos casos, se indica la dirección del núcleo (p. e., se diferencia entre adjuntos nominales de nombre situado a la izquierda o a la derecha).

(3) REMOVE TARGET (@SUBJ) IF (0 NOM) (-1C PREP);

2 CATCG: estado del proyecto y perspectivas de futuro

Datos técnicos:

- Tamaño del léxico: aprox. 90.000 lemas
- Velocidad de procesamiento: 1800 p/s
- Plataformas: unix, linux

	DeMCat	ASCat
Precisión	0.92	0.78
Cobertura	0.98	0.96
F (a=0.5)	0.95	0.87

La relativamente baja precisión de CATCG se debe a la voluntad de conseguir un muy bajo porcentaje de error a favor de la cobertura, es decir, de utilizar sólo reglas muy fiables. Con el formalismo CG, y con esta aproximación, calculamos que el techo está en un 90%-95% de precisión global. Algunas de las ambigüedades persistentes se procesarán en módulos posteriores: actualmente se está desarrollando un módulo para tratar adjunción de sintagmas preposicionales.

Otra de las líneas de investigación actuales es la de adquirir automáticamente información semántica para explotarla en todos los módulos y en aplicaciones posteriores.

3 Proyectos

CATCG es una herramienta básica que ya se está aplicando en varios proyectos en desarrollo en el seno de GLiCom:

- **BancTrad** (http://glotis.upf.es/bt/index.html) es una interfaz de búsqueda para corpus paralelos anotados con información lingüística y extralingüística. Sus usuarios potenciales son estudiantes de traducción, traductores, lingüistas y otros profesionales de la lengua. Financiación: *Programa d Innovació Docent,* UPF. Duración: 01/2000-12/2002.
- El proyecto **PrADo** (*Preparación Automatizada de Documentos*) tiene como objetivo la creación de correctores gramaticales para español y catalán, con especial hincapié en las interferencias entre estas dos lenguas y entre las mismas y el inglés. Financiación: proyecto TIC2000-1681-C02-01 del MCYT. Duración: 01/2001-12/2003.
- En el proyecto **ALLES** (*Automatic Long-distance Language Education System*) se creará una plataforma de teleenseñanza para la adquisición de las competencias lingüísticas oral y escrita (tanto en la producción como en la comprensión) dirigida a aprendices de segundas lenguas en el ámbito de la economía. Financiación: proyecto IST-2001-34246 del Vth RTD FrameWork Programme de la UE.Duración: 06/2002-06/2005.

Bibliografía

Alsina, A. *et al*.2002. CATCG: a general purpose parsing tool applied, en *Proceedings of LREC2002*, Las Palmas

Badia, T., À. Egea, T. Tuells. 1997. CATMORF: Multi-two level steps for Catalan morphology. *Demo Proceedings of the Conference on Applied Natural Language Processing.* Washington

Badia, T. *et al*. 2002. BancTrad: a web interface for integrated access to parallel annotated corpora, en *Proceedings of the LREC2002 workshop on Language Resources for Translation Work and Research*, Las Palmas, 28 mayo 2002

Karlsson, F. et al. 1995. *Constraint Grammar: a Language-Independent Formalism for Parsing Unrestricted Text.* Berlin/New York: Mouton De Gruyter