Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 455-456 recibido 20-05-2005; aceptado 01-06-2005

TXALA un analizador libre de dependencias para el castellano*

Jordi Atserias, Eli Comelles

TALP Research Center
Universitat Politècnica de Catalunya
{batalla,comelles}@lsi.upc.edu

Aingeru Mayor

IXA Group
Euskalerriko Unibersitatea
aingeru@si.ehu.es

Resumen: Esta demostración presenta la primera versión de Txala, un analizador de dependencias para el castellano desarrollado bajo licencia LGPL. Este analizador se enmarca dentro de la generación de una plataforma de software libre para la traducción. La carencia de este tipo de analizadores sintácticos para el castellano, hace que esta sea una herramienta necesaria para el progreso del PLN en castellano.
Palabras clave: Sintaxis, Analizador Sintáctico de Dependencias, Software Libre

Abstract: In this demo we present the first version of Txala, a dependency parser for Spanish developed under LGPL license. This parser is framed in the development of a free-software platform for Machine Translation. Due to the lack of this kind of syntactic parsers for Spanish, this tool is essential for the development of NLP in Spanish.
Keywords: Syntax, Dependency Parser, Free Software

1. Introducción

En esta demostración presentamos un analizador de dependencias para el castellano bajo licencia GNU Lesser General Public License (LGPL). Este analizador libre de dependencias se incluye dentro de un proyecto de Traducción Automática de código abierto para las lenguas del Estado español.

Nuestro analizador forma parte del módulo de análisis del sistema de traducción automática de transferencia profunda para castellano-euskara. El hecho de usar un analizador de dependencias se debe a la gran divergencia lingüística que existe entre el castellano y el euskara.

Cabe destacar que encontramos diversos trabajos sobre el análisis de dependencias del inglés y su implementación (Mel'cuk, 1988; By, 2004), sin embargo esto no es tan común en una lengua de orden variable como el castellano. De hecho, hay muy pocos analizadores de dependencias del castellano, entre ellos DILUCT (Calvo, Gelbuck, y Kilgariff, 2005) y Connexor[1]. La falta de analizadores de dependencias para el castellano y más aun bajo licencia LGPL, hacen de este prototipo un recurso valioso para el PLN. Durante la presentación describiremos la arquitectura del analizador y finalmente esbozaremos las líneas de trabajo futuro.

2. TXALA parser

Para obtener este analizador de dependencias decidimos extender un analizador sintáctico ya existente en Freeling (Carreras et al., 2004). Para transformar un árbol de análisis a dependencias es necesario determinar qué nodos actúan como núcleos en cada nivel del árbol. Para obtener estos núcleos es necesario adaptar tanto la gramática como el propio analizador. La adaptación de la gramática consistirá en marcar en cada regla cual es el elemento que actúa como *núcleo*. Por ejemplo, en la regla: *sn* $\Longrightarrow$ *espec-ms, grup-nom* que produce un sintagma nominal (*sn*) a partir de un artículo (*espec-ms*) y un nombre (*grup-nom*), marcaríamos el nombre *grup-nom* como el núcleo de esta regla, de manera que al extraer las dependencias del árbol, el artículo dependería del nombre. Por otra parte, la adaptación del analizador ha consistido en la extensión del formalismo de la gramática para permitir marcar núcleos y la inclusión de esta información al árbol de

* Esta investigación ha sido parcialmente financiada por el Ministerio de Industria, Turismo y Comercio PROFIT FIT-340101-2004-3

[1] http://www.connexor.com/

análisis resultante.

No siempre es posible obtener un árbol de análisis completo a partir de la gramática, La idea final detrás de este proceso es construir un módulo que a partir de un análisis sintáctico basado en *chunks* complete el análisis de la frase. Los fragmentos de análisis obtenidos de la etapa anterior se combinan mediante reglas sintácticas básicas de izquierda a derecha. La Figura 1 muestra dos reglas, la primera combina un *sn* a la izquierda de un *grup-verb*, donde el + indica la posición del núcleo, mientras que la segunda combina un *grup-verb* a la izquierda de un *sn*.

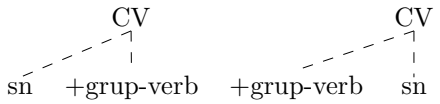

Figura 1: Ejemplos de Reglas

Por ejemplo, a partir de los *chunks* obtenidos para la frase *Mi gato, Txistu, come pescado* (ver Figura 2) y en pasos sucesivos, se combina el primer *chunk* ($sn_{mi-gato}$) con el *chunk* a su derecha ($F_{,}$), quedando $sn_{mi-gato}$ como núcleo y formando un nuevo árbol que a su vez se combinará con el siguiente *chunk*, y así sucesivamente hasta completar la frase.

Una vez construido un árbol sintáctico completo se extraen las dependencias mediante el método descrito en (Haro, 2000).

3. Trabajo Futuro

Además de los errores en el preproceso (tokenización, PoS tagging), al basarse este primer prototipo sólo en criterios sintácticos, las principales causas de errores son las decisiones de *PP-atachment*. Como trabajo futuro, tedremos que construir un módulo que combine criterios sintácticos y semánticos para completar el árbol de análisis a partir de los *chunks* obtenidos.

También será necesario un estudio más profundo sobre la estructura de las dependencias de ciertas construcciones sintácticas (p.e. frases subordinadas de relativo o la coordinación).

Por el momento no hemos tenido ocasión de evaluar exhaustivamente este prototipo, pero la construcción de un tree bank de dependencias a partir del corpus 3LB (Gelbukh, Calvo, y Torres, 2005) crea el marco necesario para una futura evaluación.

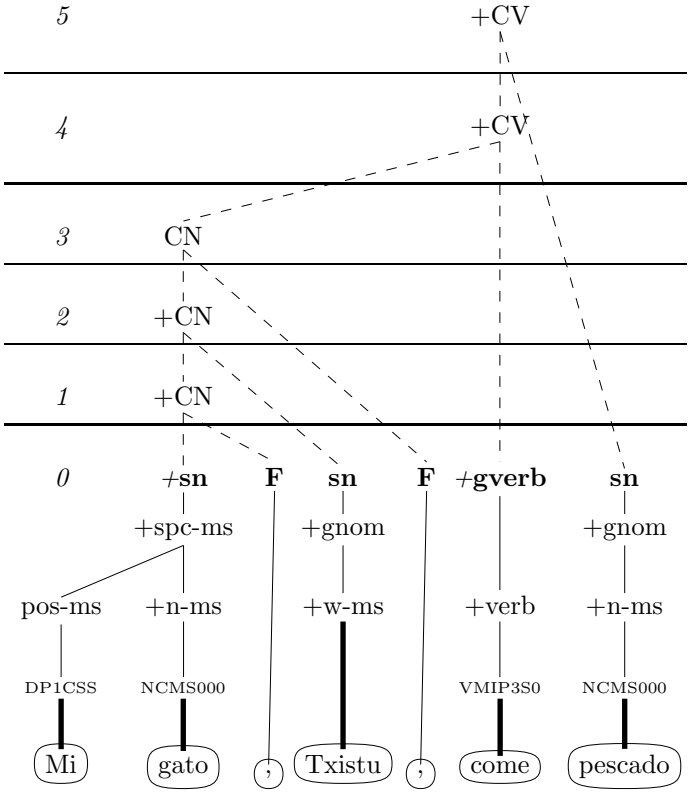

Figura 2: Árbol de análisis

Bibliografía

By, Tomas. 2004. English dependency grammar. En *RADG 2004*.

Calvo, Hiram, Alexander Gelbuck, y Adam Kilgariff. 2005. Distributional thesaurus versus wordnet. En *Proceedings of CICLING 2005*.

Carreras, Xavier, Isaac Chao, Lluís Padró, y Muntsa Padró. 2004. Freeling: An opensource suite of language analyzers. En *Proceedings of the 4th International Conference LREC'04*.

Gelbukh, A., H. Calvo, y S. Torres. 2005. Transforming a constituency treebank into a dependency treebank. En *Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'2005)*.

Haro, Sofia. 2000. *Analizador sintáctico avanzado dirigido por el diccionario de patrones de manejo sintáctico*. Ph.D. tesis, CIC-IPN.

Mel'cuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. Stte University of New York Press.