Adquisición de recursos básicos de lingüística computacional del gallego para aplicaciones informáticas de tecnología lingüística

Luz Castro Pena
Ángel López López
José Ramom Pichel Campos
imaxin|software
rúa Entremurallas, 5
15702 Compostela
imaxin@imaxin.com

José Luis Aguirre Moreno
Alberto Álvarez Lugrís
Xavier Gómez Guinovart
Elena Sacau Fontenla
Lara Santos Suárez
Seminario de Lingüística Informática
Universidade de Vigo
sli@uvigo.es

Resumen: Este trabajo presenta las características principales del proyecto Empresa-Universidad "Estudio y adquisición de recursos básicos de lingüística computacional del gallego para la elaboración y mejora de aplicaciones informáticas de tecnología lingüística" desarrollado por imaxin|software y el Seminario de Lingüística Informática (SLI) de la Universidade de Vigo.
Palabras clave: recursos lingüísticos, anotación morfosintáctica, verificación gramatical, traducción automática, lengua gallega

Abstract: This work presents the main features of the project "Acquisition of basic resources in Galician computational linguistics for language engineering" (Xunta de Galicia, 2001-2003, ref. PGIDT01TICC06E) led by imaxin|software and the Computational Linguistics Group (SLI) of the University of Vigo.
Keywords: language resources, morphosyntactic tagging, grammar checking, machine translation, Galician language

1 Datos del proyecto

1.1 Título del proyecto

Estudio y adquisición de recursos básicos de lingüística computacional del gallego para la elaboración y mejora de aplicaciones informáticas de tecnología lingüística.

1.2 Institución financiera

Secretaría Xeral de Investigación e Desenvolvemento, Xunta de Galicia, 2001-2004 (ref. PGIDT01TICC06E).

1.3 Grupos participantes

Proyecto Empresa-Universidad. Empresa: imaxin|software, Compostela. Universidad: Seminario de Lingüística Informática (Universidade de Vigo).

1.4 Investigadores principales

- José Ramom Pichel Campos (jramompichel@imaxin.com). imaxin|software, rúa Entremurallas, 5, 2º. 15702 Compostela. Telf. 981 554068.
- Xavier Gómez Guinovart (xgg@uvigo.es): Seminario de Lingüística Informática (Universidade de Vigo). Telf. 986 813858.

2 Objetivos

Este proyecto de investigación Universidad-Empresa se orienta a la adquisición de recursos básicos para la lengua gallega, con vistas a su incorporación en diversas aplicaciones informáticas de tecnología lingüística. En esta ocasión, se estudiaron las posibilidades de utilización de estos recursos lingüístico-computacionales en dos aplicaciones informáticas para el gallego que poseen un alto nivel de incidencia social: por un lado, la

verificación gramatical y estilística, en el marco del procesamiento de textos para Office XP (o productos superiores de Microsoft); y por otra, la traducción automática trilingüe entre las lenguas gallega, española y portuguesa.

3 Metodología

En primer lugar, se elaboró un etiquetador morfosintáctico automático, lo que supuso el diseño previo de un etiquetario (*tagset*) lingüísticamente motivado para el gallego y la implementación de este etiquetario en un léxico computacional del gallego (Aguirre et al. 2002, 2003). La elaboración de este etiquetario se realizó de acuerdo con las directrices estándares de la Unión Europea elaboradas por EAGLES (Monachini y Calzolari, 1996). La realización del léxico computacional del gallego implicó la elaboración de una descripción morfológica formal completa y computacionalmente eficiente de las características y clases flexivas del gallego. La técnica de descripción empleada consistió en una morfología de estados finitos que define las equivalencias entre le(xe)mas y palabras flexionadas del gallego, permitiendo al mismo tiempo la generación y análisis morfológico y la identificación de las etiquetas morfosintácticas y de los le(xe)mas correspondientes a cada palabra generada o analizada por el sistema. El proceso de etiquetación emplea un algoritmo probabilístico de desambiguación y produce textos etiquetados en el estándar XML.

Como parte del análisis de la implementación de un sistema de verificación lingüística automática para el gallego, se realizó un amplio estudio del error gramatical y de los problemas estilísticos de la lengua gallega escrita a través del vaciado de obras normativizadoras. Los errores y problemas analizados se codificaron para su procesamiento automático en la herramienta de verificación gramatical (*grammar checker*) y estilística (*style checker*) en desarrollo para Office.

En cuanto al estudio del empleo de recursos lingüísticos en aplicaciones de traducción automática entre el gallego, español y portugués, se constituyó un corpus trilingüe de estas lenguas y se establecieron las correspondencias entre el etiquetario diseñado para el gallego y el etiquetario intermedio (*intermediate tagset*) propuesto por EAGLES como representación estándar y lingüísticamente neutral de la información lingüística codificada en las etiquetas. La incorporación al proyecto de este etiquetario intermedio nos permite aprovechar la correspondencia entre la información gramatical codificada para el gallego en el etiquetario del SLI y la que se codifica en formato estándar de EAGLES en aplicaciones de otras lenguas (concretamente, en este proyecto, en aplicaciones para la lengua castellana y para la lengua portuguesa). El objetivo del alineamiento a distintos niveles del corpus trilingüe etiquetado morfosintácticamente es la extracción de información traductológica para aplicaciones de traducción automática basada en corpus.

Bibliografía

Aguirre, J.L., A. Álvarez Lugrís y X. Gómez Guinovart. 2002. Etiquetario morfosintáctico del SLI para corpus de lengua gallega: aplicación al corpus paralelo TECTRA. *Procesamiento del Lenguaje Natural*, 28: 23-34.

Aguirre, J.L., A. Álvarez Lugrís y X. Gómez Guinovart. 2003. Aplicación do etiquetario morfosintáctico do SLI ó corpus de traduccións TECTRA. *Viceversa: Revista Galega de Traducción*, 7.

Aguirre Moreno, J.L., N. Andión y X. Gómez Guinovart. 2001. Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega. *Procesamiento del Lenguaje Natural*, 27: 13-19.

Aguirre, J.L., A. Álvarez Lugrís, I. Bragado, L. Castro, X. Gómez Guinovart, S. González Lopo, A. López López, J.R. Pichel, E. Sacau y L. Santos. 2003. Alinhamento e etiquetagem de corpora paralelos no CLUVI (Corpus Lingüístico da Universidade de Vigo). En Almeida, J.J. (ed.), *Actas do Workshop CP3A, Corpora Paralelos: Aplicações e Algoritmos Associados*, Universidade de Braga (Portugal).

Monachini, M. y N. Calzolari (coords.). 1996. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. EAGLES Guidelines.