

# ANOTACIÓN AUTOMÁTICA DE TEXTOS DIACRÓNICOS DEL ESPAÑOL

Cristina Sánchez Marco  
cristinadesu@gmail.com  
*Universidad de Zaragoza*

Josep Maria Fontana  
josepm.fontana@upf.edu  
*Universitat Pompeu Fabra*

Judith Domínguez  
Mañosa  
*Universitat Pompeu Fabra*

Para poder realizar una investigación empírica sobre la evolución histórica de una lengua es fundamental tener un conjunto de textos en los que analizar el fenómeno. Actualmente, el lingüista o filólogo puede utilizar herramientas computacionales para identificar y extraer los casos de los textos con mayor rapidez, así como para realizar análisis cuantitativos de los datos de manera automática. En el caso del español, sin embargo, los corpus diacrónicos representativos accesibles en la Web para la comunidad investigadora —como son el Corpus del español, desarrollado por Mark Davies<sup>1</sup>, y el CORDE de la Real Academia Española<sup>2</sup>— están escasamente anotados con información lingüística y la interfaz de usuario solo permite hacer búsquedas simples, lo que limita enormemente su uso para estudios lingüísticos. Además, carecen de la posibilidad de visualizar la información paleográfica del documento original, muy valiosa para los estudios filológicos. Por todo ello, hemos iniciado un proceso de compilación y anotación automática de textos diacrónicos en español. A continuación describimos el método desarrollado, así como el estado del proyecto.

Uno de los principales obstáculos con los que nos encontramos cuando intentamos enriquecer un corpus diacrónico con anotaciones lingüísticas de manera automática es la enorme variación ortográfica que caracteriza a este tipo de textos. Esto hace que en algunos de los proyectos de anotación de corpus diacrónicos se haya optado por la anotación manual o de los mismos —por ejemplo, el Corpus do Português, desarrollado por Mark Davies y Michael Ferreira<sup>3</sup>, y el Tycho Brahe Parsed Corpus of Historical Portuguese, de Galves y Britto (2003)—, con la consiguiente inversión de tiempo y recursos. En el proyecto que describiremos en esta presentación hemos querido explorar una estrategia alternativa: aplicar herramientas existentes para el procesamiento del español contemporáneo a documentos históricos a los que se les ha añadido información sobre las relaciones entre las grafías antiguas y sus equivalentes en español contemporáneo. En concreto explicaremos cómo hemos utilizado un etiquetador automático de base estadística que asigna un lema y una categoría a cada palabra para etiquetar morfosintácticamente un texto que ha sido previamente estandarizado de manera automática. Además, mediante el uso del lenguaje de marcado XML y siguiendo los estándares del TEI hemos podido preservar toda la información del documento original, incluida la información paleográfica añadida por los editores del texto. La estrategia desarrollada es en cierta manera un híbrido de las aproximaciones desarrolladas para algunos corpus diacrónicos del inglés, como el *Helsinki Corpus of English Texts*, y algunas herramientas de preprocesamiento que facilitan la creación de un corpus en el que las variantes ortográficas están modernizadas (Baron y Rayson 2008). Con el enfoque que proponemos, los resultados

---

<sup>1</sup> <http://www.corpusdelespanol.org>

<sup>2</sup> <http://www.rae.es>

<sup>3</sup> <http://www.corpusdoportugues.org>

alcanzan el 90% de precisión en etiquetado morfológico y alrededor del 85% para la asignación de lema.

Aproximaciones como la presentada en este trabajo son muy relevantes para la creación de recursos lingüísticos para el estudio de la historia del español, ya que la anotación lingüística facilita enormemente las aproximaciones empíricas al análisis lingüístico y el procesamiento automático a menudo es la única opción viable para obtener dicha anotación a no ser que se disponga de abundantes recursos económicos y humanos.

## Referencias

- BARON, A. and Rayson, P. 2008: "VARD 2: A tool for dealing with spelling variation in historical corpora", en *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, 22 May 2008.
- GALVES, Charlotte y Britto, Helena (2003): "A Construção do Corpus Anotado do Português Histórico Tycho Brahe: o sistema de anotação morfológica", en [http://www.ime.usp.br/~tycho/participants/c\\_galves/galves\\_e\\_britto.htm](http://www.ime.usp.br/~tycho/participants/c_galves/galves_e_britto.htm)