Limitaciones en el uso de corpus diacrónicos del español.
Nuevas aportaciones desde el proyecto de investigación *Post Scriptum*.*

Existen en la actualidad varios corpus diacrónicos en formato electrónico. En el caso del español, los dos grandes corpus en línea para el estudio de la diacronía son el Corpus Diacrónico del Español (CORDE) y el Corpus del Español (CE). El volumen de texto recopilado –250 y 100 millones de palabras, respectivamente– proporciona una fuente de datostodavía hoy inexistentepara otras lenguas,lo queconvierte estos corpus en herramientas de referencia para la investigación en Lingüística Histórica.

Asumidoel indiscutible valor de talesrecursos, no es menos cierto que estos macrocorpus presentanalgunas limitaciones. En primer lugar, hay que destacarla falta de anotación lingüística detallada. A día de hoy, el CORDE todavía no está lematizado, lo que restringe considerablemente el tipo de información recuperable. El CE sí permite búsquedas por lema, peroresulta inadecuadopara el estudio de fenómenos gramaticales que demanda una anotación más compleja (Nieuwenhuijsen 2009; García Salido y Vázquez Rozas 2012).En segundo lugar, la información extratextual de las obras recopiladas se ciñe a los aspectos más generales: marco cronológico, procedencia geográfica y género textual. Esto impide al investigador el controlsobre otros factores que influyen en las opciones lingüísticas de quien escribe, como son la caracterización social de autor y destinatario, su procedencia dialectal o la intención comunicativadeltexto(Enrique-Arias 2012).Finalmente, existe una tercera restricción que es consustancial a cualquier conjunto de datos no contemporáneo: la imposibilidad de incluir fuentes orales.

En este trabajo se profundizará en las limitaciones que presentan actualmente los corpus diacrónicos sobre el español y se dará a conocer el proyecto de investigación Post Scriptum, una nueva herramienta digital que pretende dar solución a los problemas anteriormente citados. El proyecto Post Scriptum, que se desarrolla en la Universidad de Lisboa, tiene entre sus objetivos la creación de un corpus compuesto por 7000 cartas privadas escritas en español (3500) y portugués(3500) durante la Edad Moderna. Estas epístolas –en su mayoría inéditas– fueron conservadas como prueba instrumental dentro de procesos de tribunales civiles y religiosos. Con frecuencia, la información incluída en los procesos judiciales permite obtener datos biográficos sobre los participantes de las cartas procesadas.

La parte española del corpus está siendo lematizada y etiquetada morfosintácticamente con ayuda del anotador Freeling 3.0 (Padró y Stalinovsky2012).Además, para cada carta se proporciona información pragmática (tipo enunciativo) e información contextual (contexto situacional). Todo ello permitirá al investigador la consulta de diferentes aspectoslingüísticos,posibilitandoasimismola aplicaciónde búsquedas cruzadas (p. ej.clíticos de segunda persona en función de objeto directo en cartas de amenaza escritas en el siglo XVII por autores de baja condición social).

Finalmente, en el ámbito de la Lingüística Diacrónica la naturaleza dialógica de estos documentos privados permite compensar, en su justa medida, la carencia de fuentes orales. Un repertorio de cartas propias de contextos informales, producidas por manos poco instruidas y escritas casi como si fuesen habladas constituye un recurso extraordinario para el estudio fonológico, morfológico y sintáctico de un determinado período histórico.

Bibliografía:

Davies, Mark (2002-):*Corpus del Español: 100 million words, 1200s-1900s*. Disponible online en http://www.corpusdelespanol.org

Enrique-Arias, Andrés (2012):"Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad". *Scriptum digital1*, pp. 85–106

García Salido, Marcos y Victoria Vázquez Rozas (2012): "Los corpus diacrónicos como instrumento para el estudio del origen y distribución de la concordancia de objeto en español." *Scriptum Digital* 1, pp. 67–84

Niewenhuijsen, Dorien (2009): "El rastreo del desarrollo de algunos pronombres personales en español: (im)posibilidades de los corpus diacrónicos digitales". En Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Frankfurt am Main: Iberoamericana/Vervuert, pp. 365-384

Padró, L. & Stanilovsky, E. (2012):"FreeLing 3.0: Towards Wider Multilinguality". *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* ELRA. Estambul, Turquía. 2012.

REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español.*http://www.rae.es

* Funded by the European Research Council, ERC Advanced Grant 2011, GA 295562.