***Post Scriptum: Archivo Digital de Escritura Cotidiana* ***

Rita Marquilhas (rita.marquilhas@gmail.com),
Ana Luísa Costa, Clara Pinto, Fernanda Pratas, Gael Vaamonde
(CLUL, Universidade de Lisboa)

1. Resumen del proyecto

El proyecto *Post Scriptum: Archivo Digital de Escritura Cotidiana* (P.S.) tiene como finalidad recoger y publicar cartas privadas portuguesas y españolas escritas durante la Edad Moderna (desde el s. XVI hasta inicios del s. XIX) por personas pertenecientes a diferentes estratos sociales. En la comunicación, se explicará la metodología utilizada en la edición digital de los documentos para su disponibilidad en línea; además, se discutirán cuestiones relacionadas con la modernización automática de los textos y con su posterior etiquetación morfológica y sintáctica.

A raíz de la relevancia que ha ido adquiriendo el estudio diacrónico de textos no literarios, en las últimas décadas ha cobrado especial interés la elaboración de corpus formados por textos de naturaleza epistolar (Daybell, J., 2012; Dossena, M. y Gabriella Camiciotti, 2012, entre otros). La edición digital en línea de *Post Scriptum* ofrece un conjunto heterogéneo de cartas escritas en diferentes contextos sociales, que obedecen a situaciones comunicativas variadas. 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. Por un lado, la espontaneidad en las interacciones que se generan a través de la correspondencia privada puede servir como una ventana al discurso cotidiano; por otro lado, 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. Finalmente, los datos biográficos de individuos anónimos, así como sus formas de vida y sus interrelaciones sociales suponen un valor de interés indiscutible, tanto desde la perspectiva histórica y cultural como desde la historiografía moderna.

2. Tarea paleográfica y edición digital

Se darán a conocer los criterios adoptados para la elaboración de la edición electrónica, prestando especial atención a los argumentos que sustentan la idea de que este proceso de edición no sólo garantiza el rigor filológico, sino que lo enriquece gracias a los medios tecnológicos disponibles en el proyecto. Para la edición de las cartas en formato digital, se han adoptado las normas de codificación propuestas por el proyecto *Digital Archive of Letters in Flanders* (DALF) para textos epistolares, que se basan a su vez en las directrices recogidas por el consorcio *Text Encoding Initiative* (TEI) para fuentes primarias. La codificación se ha realizado mediante lenguaje XML. Una vez finalizado, cada documento XML puede ser visto como una cartografía de su correspondiente manuscrito: junto a la transcripción y el facsímile, cada carta va acompañada de metadatos, palabras clave, información histórica y contextual, traducción al inglés y una versión con la ortografía y la puntuación normalizadas. En definitiva, lo que se ofrece es un conjunto de datos que facilita la investigación en línea para múltiples campos de estudio, con diferentes herramientas electrónicas.

3. Investigación lingüística

El proyecto P.S. se propone enriquecer la investigación en Lingüística Diacrónica, beneficiándose para ello de los avances que proporcionan actualmente la Lingüística de Corpus y la Lingüística Computacional. Conscientes de que la normalización ortográfica permite obtener mejores resultados en tareas de anotación morfológica y léxica, los textos portugueses se están normalizando semi-automáticamente mediante la herramienta VARD 2 para el portugués (Hendrickx, I. Y Marquilhas, R., 2011). Respecto al etiquetado gramatical, el conjunto de cartas portuguesas se está anotando con la herramienta Edictor, desarrollada por el proyecto Tycho Brahe para el portugués (Galves, C. y Britto, H., 2002); para la parte española del corpus, se está utilizando el anotador de FreeLing 3.0 (Padró, L. & Stalinovsky, E., 2012).

El sistema utilizado para el etiquetado gramatical del corpus portugués es compatible con los *parsers* elegidos, los cuales se basan en el sistema de anotación sintáctica desarrollado por el *Penn Corpora of Historical English* (Kroch, T., Santorini, B.& Diertani, A., 2010). Para la anotación sintáctica del corpus español, se ha recurrido de nuevo a la herramienta FreeLing 3.0. Puesto que el sistema de etiquetas utilizado por el parser de FreeLing es más completo que el utilizado por el *Penn Corpora of Historical English*, siempre es posible convertir automáticamente las etiquetas de un formato a otro.

4. Observaciones finales

La edición digital y la anotación de corpus en este proyecto implica, más que un proceso ascendente (de la transcripción paleográfica a la anotación sintáctica), una relación dinámica entre diferentes niveles de actuación.

Actualmente, se puede utilizar la edición digital de 2000 cartas privadas portuguesas (más de 600.000 palabras), de los siglos XVI al XIX (http://alfclul.clul.ul.pt/cards-fly).

5. Bibliografía

DALF, *Guidelines for the description and encoding of Modern correspondence material* (http://ctb.kantl.be/project/dalf/).

Daybell, J. (2012). *The material letter in Early Modern England. Manuscript letters and the cultura and practices of letter writing, 1512-1635.* Hampshire: Palgrave Macmillan. Dossena, M. & Camiciotti, G. (2012). *Letter writing in late modern Europe*. Amsterdam: John Benjamins.

FreeLing (http://nlp.lsi.upc.edu/freeling/).

Galves, C. & Britto, H. (2002). *The Tycho Brahe Corpus of Historical Portuguese*. Department of Linguistics, University of Campinas. Online publication, first edition (http://www.tycho.iel.unicamp.br/~tycho/).

Hendrickx, I. & Marquilhas, R. (2011) From old texts to modern spellings: an experiment in automatic normalisation, *Journal for Language Technology and Computational Linguistics* 26, n.º 2, 65-76.

Kroch, A., Santorini, B. & Diertani, A. (2010). *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition (http://www.ling.upenn.edu/hist-corpora/).

Lopes, et al. *Corpus Compartilhado Diacrônico: cartas pessoais brasileiras* (http://www.letras.ufrj.br/laborhistorico/).

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

TEI, *Text Encoding Initiative* (http://www.tei-c.org/index.xml

* Proyecto financiado por el European Research Council, ERC Advanced Grant 2011, GA 295562.