The Málaga Corpus of Late Modern English Scientific Prose

Javier Calle-Martín
University of Málaga

Keywords: Corpus compilation, Late Modern English, Málaga Corpus, Tagging

Abstract

The Málaga Corpus of Early English Scientific Prose is a collection of English vernacular medical writing, consisting of three diachronically divided components, i.e. The Málaga Corpus of Late Middle English Scientific Prose (1350-1500); The Málaga Corpus of Early Modern English Scientific Prose (1500-1700); and The Málaga Corpus of Late Modern English Scientific Prose (1700-1900). The three components have been purposely designed so as to contain evidence from the three text types of medical writing in English, that is, theoretical treatises, surgical treatises and recipe collections. In itself, the corpus stems from actual linguistic evidence of the period, both handwritten and printed, standing out as the ideal input for diachronic linguistic research at the levels of spelling, morpho-syntax and lexis.

The present paper is particularly concerned with the third component of the corpus, *The Málaga Corpus of Late Modern English Scientific Prose* (1700-1900), which has been recently published and made available in the project's webpage (https://latemodernmss.uma.es). In its current form, the corpus amounts to 2.5 million words, of which 1.5 million belong to the 18th century and the other million to the 19th century. The corpus is offered in three different formats, that is, the plain text version, the modernised version and the tagged version. The CQP-web version is also available for online use (https://latemodernmss.uma.es/cqpweb/). The present paper first describes the rationale of the corpus considering the typology of texts, their chronology, the text types and authorship. Second, the paper delves into the process of compilation, which is a sequential process consisting of a) modernisation by means of VaRD (Variant Detector) and b) automatic tagging by means of CLAWS (Constituent Likelihood Automatic Word-tagging System). The paper closes with a brief demonstration of the corpus potential using the CQP-web version.

References

Calle-Martín, Javier, Miriam Criado-Peña, Verónica Hernández, Sinéad Linehan-Gómez and Juan Lorente-Sánchez. 2016. The Málaga Corpus of Late Modern English Scientific Prose (MCLModESP). Málaga: University of Málaga. Available from https://latemodern.uma.es.