The Corpus Mallorca.
Advantages of a dual edition system for a historical corpus

Andrés Enrique-Arias

Universitat de les Illes Balears

The Corpus Mallorca (www.corpusmallorca.es) is an online textual database for the historical study of contact phenomena in the Spanish used by Catalan-dominant bilinguals in the island of Mallorca. To this end, the corpus is composed of documents written as close as possible to the vernacular, such as personal letters and court testimonies, composed by Catalan–Spanish bilinguals. In its present state the corpus contains some 600 documents (approximately half a million words of text) dated between 1670-1909. This effort is part of the CHARTA network (www.charta.es), a consortium of research groups that compile textual corpora made up of historical documents issued in different parts of the Spanish-speaking world.

The main challenge in preparing an electronic searchable corpus of vernacular texts written by semi-learned individuals is presenting an intelligible text, clean of graphic variation as to not obscure searches, and at the same time maintaining all the linguistic features of potential interest for linguistic research (Baron & Rayson 2008). None of the alternatives offered to corpus developers are completely satisfactory: intervening on the text can make it more intelligible but it involves loss of relevant information about the graphic configuration of the original and, at the same time, any attempt to maintain the scriptural idiosincrasies of the manuscript entails a loss of clarity along with added difficulties for text lemmatization and POS tagging (Rayson et al. 2007; Taulé et al. 2015).

In this presentation we will review the main philological and technical issues that have arisen in the creation of the Corpus Mallorca, with particular attention to the paleography vs. normalization dilemma. Following the guidelines set forth for the CHARTA network (cf. Sánchez-Prieto Borja 2011) our approach is based on a dual edition system. For each of the texts in the corpus, two versions are prepared: a paleographic transcription that preserves the graphic options of the original, and a normalized version in which texts are edited to reflect conventional spelling and punctuation. Users can then choose to search on any one of the two versions or both, but then when visualizing the results both versions appear arranged in parallel. In addition, users can access the digital images of the original manuscript.

As will be illustrated with various case studies of lexical, morphosyntactic and phonetic searches, this multiple presentation of the texts yields some practical advantages: as the dual edition includes a paleographic version, the interventions on the normalized version do not entail irreversible loss of information; at the same time, the multiple presentation affords maximal flexibility as it facilitates obtaining data for different levels of analysis (i.e. the paleographic transcription for locating grapho-phonetic features, and the normalized version for morphosyntactic and lexical phenomena). Finally, one additional advantage of the double edition is that there is no need for complex tagging in order to link alternate spellings of the same word to a single lemma as the lemmatization is done over the normalized version.

REFERENCES

Baron, A. & P. Rayson (2008). “VARD2: a tool for dealing with spelling variation in historical corpora”. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, 22 May 2008. On line at: <http://eprints.lancs.ac.uk/41666/1/BaronRaysonAston2008.pdf>

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). “Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora”. In *Proceedings of Corpus Linguistics 2007*, July 27-30, University of Birmingham, UK.

Sánchez-Prieto Borja, P. (2011). *La edición de textos españoles medievales y clásicos. Criterios de presentación gráfica*. San Millán de la Cogolla: Cilengua.

Taulé, Mariona, M. A. Martí, A. Bies, M. Nofre, A. Garí, Z. Song, S. M. Strassel & J. Ellis (2015). “Spanish Treebank Annotation of Informal Non-standard Web Text”. *ICWE Workshops*, 15-27