A corpus of seventeenth-century English news reportage: construction, encoding and applications

Dawn Archer, Andrew Hardie, Tony McEnery & Scott Piao
Dept. Linguistics, Lancaster University

This poster describes a 750 thousand-word corpus of news reportage from the English Civil War, currently under construction at Lancaster, using texts drawn from the Thomason Tracts. This collection (held at the British Library) is unique in containing the greater part of what was published in London in the period 1640-1661.

We are in the process of transcribing the periodical newsbooks published between December 1653 and May 1654 to an SGML-based format. The fairly light markup initially used by the transcribers is similar to HTML, allowing the texts subsequently to be mapped automatically to both a TEI-compliant SGML/XML format and a web-compatible HTML format, facilitating the widest possible potential re-use of the corpus.

Simultaneous with the development of the corpus, a number of linguistic/historical issues have been investigated using the transcribed newsbooks. These include an examination of the complicated nature of text re-use in the press at this period, and an inquiry into the presentation of women in different newsbooks.