Manual to accompany
The Enhanced Shakespearean Corpus: EEBO-TCP Segment (ESC: EEBO)
Lancaster University

1. Design of the corpus

The ESC: EEBO-TCP Segment was compiled as broad corpus to assist in contextualising Shakespeare's plays for the Encyclopedia of Shakespeare's Language Project (AHRC grant reference AH/N002415/1). The compilation was undertaken by Sean Murphy and Andrew Hardie (Lancaster University) with input from other project colleagues between 2016 and 2019. The ESC: EEBO comprises some 380 million words spanning the 80-year period 1560-1639 and incorporates diverse genres. More detail on the compilation of the corpus, including annotation, can be found in Murphy (2019).

2. Source texts used for the corpus

The source texts of the ESC: EEBO corpus were all obtained from the Early English Books Online - Text Creation Partnership (EEBO-TCP); see further http://www.textcreationpartnership.org/tcp-eebo/.

3. Genre metadata: The genre classification of EEBO texts for the period 1560-1639

A key feature of the ESC: EEBO is that a genre classification scheme, adopting a prototype approach, has been applied its 5,900 texts. This classification is shown in the table below. Note that we removed any texts duplicated in ESC: Folio or ESC: Comp. We acknowledge that a degree of fuzziness and overlap amongst categories remains. Whilst some were relatively easy to separate out and place, others had rather mixed contents and / or membership claims to multiple superordinate categories. In such cases, we made a judgement about best fit.

Style	**Domain**	**Genre**	**Sub-genres**
Literary	Literature	Plays	Comedy, History, Tragedy, Masque
		Poetry, Verse & Song	Ballads, Songs
		Fiction	
		General	
Formal – Spiritual	Religion	Bible	
		Catholicism	Anti-Catholicism
		Protestantism	Church of England, Church of Scotland, Non-Conformism
		Doctrine, Theology and Governance	Heresy, Prayer, Sin and Repentance,
		General	Articles, Christians, Devotional, Epistles, Sermons, Others
Formal - Statutory	Administration	Royal	Communications and Orders, Proceedings
		Parliamentary	General, Proceedings and Reports
		Legal	Legislation and Orders, Trials and Disputes
		General	Declarations, Military, Proceedings, Speeches
Formal - Instructional	Instruction	Astronomy	
		Philosophy	
		Science	Experiments
		Mathematics	
		Medicine	Anatomy
		General	Alchemy, Almanack, Astrology and Predictions, Lecture

Informational	Information	Biography	
		Colonial	
		Essay	Admonition, Advisory, Apologia, Argumentative, Commentary On People And Places, Death, Obituaries and Epigraphs, Dialogue, Exhortation, General Lamentations
		Letters	
		Pamphlets	Analysis And Instruction, Chronology, Directory, Finance and Trade, Food and Cookery, History, Language, Travel, Treatise, London, Petitions, Reportage, Satire, Wit and Humour
		General	

4. Normalisation of spelling variation

The play-texts in the ESC: EEBO have undergone some normalisation (regularisation) of Early Modern English spelling variation. This was done in part using PHP scripts (notably to join open compounds which are now typically closed, e.g. *it self* -> *itself*), and in part using the software tool VARD 2 (see http://ucrel.lancs.ac.uk/vard/about/) in automatic mode at the 70% confidence level. The spelling normalisation is designed to improve the usability of the play-texts with corpus tools, as it improves the prospects for orthographic matching of word-forms. Note, though, that (i) some spelling variation certainly remains, and (ii) automatic spelling normalisation is subject to error.

5. Grammatical tagging

The play-texts in the ESC: EEBO have also been annotated with grammatical part-of-speech tags using a customised version of the Constituent Likelihood Automatic Word-tagging System (CLAWS; see Leech et al. 1994; http://ucrel.lancs.ac.uk/claws/). CLAWS tags are alphanumerical codes in square brackets which correspond to over 200 part of speech classifications (CLAWS tagset version 6 was used; see http://ucrel.lancs.ac.uk/claws6tags.html). For example, [JJ] denotes an adjective, [NN] a noun and [VV] a verb. Note that although the version of CLAWS used had been trained on Early Modern English play-texts (specifically, Shakespeare's plays), it has not been manually checked and there may be errors in the tagging.

6. Semantic tagging

The play-texts in the ESC: EEBO have also been annotated for semantic meaning, using the UCREL Semantic Analysis System (USAS; Rayson et al. 2004) in the Wmatrix suite of corpus linguistic software tools (Rayson 2008). USAS assigns a semantic category label (in the form of an alphanumeric tag) to each word, using a taxonomy of 232 categories of meaning grouped into 21 main semantic fields (see further http://ucrel.lancs.ac.uk/usas/). Although USAS has been successfully used for semantic analysis of historical texts, it should be noted that the USAS semantic classification system was developed for late 20th century English. Some Early Modern English words no longer in use may be unfamiliar to the tool and therefore wrongly classified. Furthermore, some word meanings may have changed between the time the plays originated and the late 20th century, again potentially resulting in errors in semantic classification.

7. Enquiries about the corpus

Enquiries about the ESC: EEBO should be directed to the Principal Investigator of the Encyclopedia of Shakespeare's Language Project, Professor Jonathan Culpeper, Linguistics and English Language Department, Lancaster University, UK, at j.culpeper@lancaster.ac.uk.

References

Baron, Alistair and Paul Rayson (2008). VARD 2: A tool for dealing with the spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, U.K.*, 22 May 2008.

Bray, Tim, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler and François Yergeau (eds.). 2008. *Extensible Markup Language (XML) 1.0.* Fifth edition. W3C Recommendation 26 November 2008. https://www.w3.org/XML/ (accessed 01.06.2019).

Murphy, Sean (2019). Shakespeare and his contemporaries: Designing a genre classification scheme for Early English Books Online 1560–1640, ***ICAME Journal 43***(1): 59-82.

EEBO-TCP = *Early English Books Online-Text Creation Partnership* (2019). http://www.textcreationpartnership.org/tcp-eebo/ (accessed 21.02.2019).

Hardie, Andrew (2014). Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal* 38:73-103.

Leech, Geoffrey, Roger Garside and Michael Bryant (1994). CLAWS 4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, Kyoto, Japan, 622—628. http://ucrel.lancs.ac.uk/papers/coling1994paper.pdf (accessed 21.02.2019).

Rayson, Paul, Archer, Dawn, Piao, Scott L. and Tony McEnery. (2004) The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 25 May 2004, Lisbon, Portugal. Paris: European Language Resources Association, pp. 7-12.