Manual to accompany
The Enhanced Shakespearean Corpus: First Folio Plus (ESC: Folio)
Lancaster University

1. Design of the corpus

The *Enhanced Shakespearean Corpus: First Folio Plus* (ESC: Folio) was compiled as the key corpus of Shakespeare's plays for the Encyclopedia of Shakespeare's Language Project (AHRC grant reference AH/N002415/1) by Jonathan Culpeper and Andrew Hardie (Lancaster University) with input from other project colleagues between 2016 and 2019. The corpus captures what might be considered Shakespeare's "canon". It includes the 36 plays published as the First Folio in 1623, plus *The Two Noble Kinsmen* and *Pericles*. More detail on the compilation of the corpus, including annotation, can be found in Culpeper et al. (2021).

2. Source texts used for the corpus

The source texts of the ESC: Folio corpus were supplied by Internet Shakespeare Editions (https://internetshakespeare.uvic.ca/), with the kind permission of the University of Victoria.

3. List of play-texts in the ESC: Folio

Play (short title)	***Abbreviation***	***Genre (tragedy, comedy, history)***	***Date of first publication***	***Date range of first production***	***Approximate date of first production***
Titus Andronicus	Tit	T	1594	1590-1592	1592
Romeo and Juliet	RJ	T	1597	1594-1595	1595
Julius Caesar	JC	T	1623	1598-1599	1599
Hamlet	Ham	T	1603	1600-1601	1601
Troilus and Cressida	TC	T	1609	1602-1603	1602
Othello	Oth	T	1622	1603-1604	1604
King Lear	KL	T	1608	1605-1606	1605
Timon of Athens	Tim	T	1623	1605-1606	1605
Antony and Cleopatra	AC	T	1623	1606-1608	1606
Macbeth	Mac	T	1623	1606	1606
Coriolanus	Cor	T	1623	1608	1608
Henry VI, Part 2	2H6	H	1594	1590-1591	1591
Henry VI, Part 3	3H6	H	1595	1591	1591
Henry VI, Part 1	1H6	H	1623	1590-1592	1592
Richard III	R3	H	1597	1591-1593	1592
Richard II	R2	H	1597	1595	1595
King John	KJ	H	1623	1596	1596
Henry IV, Part 1	1H4	H	1598	1596-1597	1597
Henry IV, Part 2	2H4	H	1600	1597-1598	1597
Henry V	H5	H	1600	1598-1599	1599
Henry VIII	H8	H	1623	1613	1613
Much Ado about Nothing	MA	C	1600	1598	1598
Two Gentlemen of Verona	TGV	C	1623	1590-1591	1590
The Taming of the Shrew	TS	C	1594	1590-1604	1592
The Comedy of Errors	CE	C	1623	1590-1594	1594
Love's Labour's Lost	LLL	C	1598	1594-1595	1595
A Midsummer Night's Dream	MND	C	1600	1595-1596	1595
The Merchant of Venice	MV	C	1600	1596-1598	1596
The Merry Wives of Windsor	MW	C	1602	1597-1598	1597
As You Like It	AYL	C	1623	1598-1600	1599
Twelfth Night	TN	C	1623	1601-1602	1601
All's Well that Ends Well	AW	C	1623	1603-1604	1603

Measure for Measure	MM	C	1623	1603-1604	1603
Pericles	Per	C	1609	1606-1608	1608
The Winter's Tale	WT	C	1623	1609-1611	1609
Cymbeline	Cym	C	1623	1608-1611	1610
The Tempest	Tem	C	1623	1611	1611
The Two Noble Kinsmen	TNK	C	1634	1613-1614	1613

*Dates of first production and first publication are from the Database of Early English Playbooks (DEEP): http://deep.sas.upenn.edu/

4. Mark-up and annotation format

The ESC: Folio texts are marked up and annotated with XML tags (see Bray et al. 2008; Hardie 2014). Each utterance is marked with an opening speaker ID tag and a close tag. One attribute of the speaker ID tag is the speaker label in its original format in the text. Original format speaker labels are often inconsistent in historical play-texts, so the speaker ID tags also contain a speaker ID label assigned by the compilers which remains consistent for that character throughout the play-text. Act and scene boundaries, stage directions, front matter, end matter and paratext, e.g. prologues and epilogues, are also marked with XML tags. Note that this kind of tagging, although widely used, may not be compatible or readable by some corpus linguistic software tools.

5. Normalisation of spelling variation

The play-texts in the ESC: Folio have undergone normalisation (regularisation) of Early Modern English spelling variation. This was done with the help of the software tool VARD 2 (see http://ucrel.lancs.ac.uk/vard/about/) running in manual (word-by-word) mode (it can on most occasions suggest regularisation options in order of likelihood, from which the human operator approves a selection). 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.

6. Grammatical tagging

The play-texts in the ESC: Folio 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. In addition, every word was manually checked for accuracy at the highest level of the tag (e.g. a word tagged NN1 and another NN2 were both checked that the initial 'N' (i.e. noun) is correct).

7. Semantic tagging

The play-texts in the ESC: Folio 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.

8. Social annotation

The play-texts in the ESC: Folio have also been annotated with XML tags for social categories. The social categories are listed in the table below. The categories relating to a character's status/social rank draw upon the scheme developed by Archer and Culpeper (2003), which reflects the nature of status in pre-industrialised Early Modern English society and the way in which Shakespeare's contemporaries wrote about it. That scheme has been slightly reworked to capture particular Shakespearean features (e.g. the category Supernatural Beings was added to account for the ghosts, gods, fairies, etc.).

Field	Feature marked	Possible values
1	Speaker(s)	Singular (s) or multiple (m)
2	Speaker ID tag	See section 4
3	Gender of speaker	Male (m), female (f), assumed male (am), assumed female (af), problematic (p)
4	Status/social rank of speaker	Monarch (0), nobility (1), gentry (2), professional (3), other middling groups (4), ordinary commoners (5), lowest groups (6), supernatural beings (7), problematic (8)

9. Enquiries about the corpus

Enquiries about the ESC: Folio 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

Archer, Dawn and Jonathan Culpeper (2003). Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics. In: Andrew Wilson, Paul Rayson and Anthony M. McEnery (eds.) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Frankfurt/Main: Peter Lang, 37-58.

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).

Culpeper, Jonathan, Hardie, Andrew, Demmen, Jane, Hughes, Jennifer and Matt Timperley (2021) Supporting the corpus-based study of Shakespeare's language: Enhancing a corpus of the First Folio. *ICAME Journal* 45(12): 37-86.

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.

Rayson, Paul (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4), 519-549.