Manual to accompany
The Enhanced Shakespearean Corpus: Comparative Plays (ESC: Comp)
Lancaster University

1. Design of the corpus

The *Enhanced Shakespearean Corpus: Comparative Plays* (ESC: Comp) was compiled as a reference corpus for Shakespeare's plays for the Encyclopedia of Shakespeare's Language Project (AHRC grant reference AH/N002415/1) by Jane Demmen and Andrew Hardie (Lancaster University) with input from other project colleagues between 2016 and 2019. The corpus is similar in size to the canon of Shakespeare's plays overall (around 1 million words), and in its proportions of comedy, history and tragedy. It contains 46 plays by 24 playwrights (22 of whom are named, two of whom are anonymous), with first production dates ranging from 1584-1626 (compared to Shakespeare's plays, written circa 1590-1613). More detail on the compilation of the corpus, including annotation, can be found in Demmen (2019).

2. Source texts used for the corpus

The source texts of the ESC: Comp were all obtained from the Early English Books Online - Text Creation Partnership (EEBO-TCP); see further http://www.textcreationpartnership.org/tcp-eebo/. Each play-text is headed with bibliographic reference details from EEBO-TCP including the Short Title Catalogue (STC) number.

3. List of play-texts in the ESC: Comparative Plays corpus

Play-text ID	Author	Title	Date of first production*	Date of first publication*	Date of edition in corpus
	Comedy				
CCCALEX	John Lyly	*Alexander and Campaspe*	c.1583	1584	1584
CCCGALL	John Lyly	*Gallathea*	1585	1592	1592
CCCFRIAR	Robert Greene	*Friar Bacon and Friar Bungay*	1589	1594	1594
CCCOLDW	George Peele	*The Old Wives Tale*	1590	1595	1595
CCCBLIND	George Chapman	*The Blind Beggar of Alexandria*	1596	1598	1598
CCCFAIR1	Thomas Heywood	*The Fair Maid of the West Part I*	1604	1631	1631
CCCANHU	George Chapman	*An Humerous Dayes Myrth*	1597	1599	1599
CCCTWOA	Henry Porter	*The Two Angry Women of Abington*	c.1598	1599	1599
CCCMUCE	Anonymous	*Mucedorus*	1590	1598	1598
CCCOLDF	Thomas Dekker	*Old Fortunatas*	1599	1600	1600
CCCCHUS	Thomas Heywood	*How a Man May Chuse*	1602	1602	1602
CCCVOLP	Ben Jonson	*Volpone*	1606	1616	1616
CCCHATER	Francis Beaumont and John Fletcher	*The Woman Hater*	1606	1607	1607
CCCMISER	George Wilkins	*The Miseries of Inforst Marriage*	1606	1607	1607
CCCKOFB	Francis Beaumont	*The Knight of the Burning Pestle*	1607	1613	1613
CCCFAITH	John Fletcher	*The Faithful Shepherdess*	1608	c.1610	1610
CCTPHILA	Francis Beaumont and John Fletcher	*Philaster*	1609	1620	1620
CCCROARI	Thomas Middleton	*The Roaring Girl*	1611	1611	1611
CCCBFAIR	Ben Jonson	*Bartholomew Fayre*	1614	1631	1631
CCCBOND	Philip Massinger	*The Bondman*	1623	1624	1624
	History				
CCHJAME	Robert Greene	*The Scottish History of James the Fourth*	c.1590	1598	1598

CCHTAMB	Christopher	*Tamburlaine Part I*	c. 1587	1590	1590
CCHEDWII	Christopher	*Edward II*	1592	1594	1594
CCHEDWA	George Peele	*The Famous Chronicle of Edward I*	1591	1593	1593
CCHPARIS	Christopher	*The Massacre at Paris*	1593	c.1594	1594
CCHALCA	George Peele	*The Battle of Alcazar*	1589	1594	1594
CCHDEAT	Anthony Munday	*The Death of Robert Earl of Huntingdon*	1598	1601	1601
CCHEDIV1	Thomas Heywood	*Edward IV Part I*	1599	1600	1600
CCHEDIV2	Thomas Heywood	*Edward IV Part II*	1599	1600	1600
CCHOLDC	Anonymous	*The Life of Sir John Oldcastle*	1599	1600	1600
CCHIFYO1	Thomas Heywood	*If You Know Not Me, You Know Nobody*	1604	1605	1605
CCHWYAT	Thomas Dekker	*Sir Thomas Wyatt*	1602	1607	1607
CCHWELS	Robert Armin	*The Valiant Welshman*	1612	1615	1615
CCHDUCH	Thomas Drue	*The Duchess of Suffolk*	1624	1631	1631
	Tragedy				
CCTSPANT	Thomas Kyd	*The Spanish Tragedy*	1587	1592	1592
CCTJEWO	Christopher	*The Jew of Malta*	1589	1633	1633
CCTFAUST	Christopher	*Dr Faustus*	1592	1604	1604
CCTDIDOC	Christopher	*Dido, Queen of Carthage*	1586	1594	1594
CCTAWK	Thomas Heywood	*A Woman Killed With Kindness*	1603	1607	1607
CCTMALC	John Marston	*The Malcontent*	1604	1604	1604
CCTSEJAN	Ben Jonson	*Sejanus*	c.1604	1604	1604
CCTMAIDT	Francis Beaumont and John Fletcher	*The Maid's Tragedy*	1610	1619	1619
CCTWHIT	John Webster	*The White Devil*	1612	1612	1612
CCTDOFM	John Webster	*The Duchess of Malfi*	1614	1623	1623
CCTCHANG	Thomas Middleton and William	*The Changeling*	1622	1653	1653
CCTWBEW	Thomas Middleton	*Women Beware Women*	1621	1657	1657

*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: Comp 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: Comp 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.

6. Grammatical tagging

The play-texts in the ESC: Comp 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.

7. Semantic tagging

The play-texts in the ESC: Comp 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: Comp 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: Comp 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).

Demmen, Jane (2019). Issues and challenges in compiling a corpus of Early Modern English plays for comparison with those of William Shakespeare. *ICAME Journal* 44: 37-68.

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.