**CRAIG — Computational**

<0 figures>

**A Computational Bibliography of Two Plays from the Shakespeare Folio**

Craig, H.

The 1623 Shakespeare Folio may well be the most important single book we have for the study of literature in English. It contains 18 plays that do not survive anywhere else, *Macbeth*, *Twelfth Night*, and *Julius Caesar* among them, and includes what are generally held to be superior texts of 18 plays by Shakespeare that had already been printed, or at least valuable alternative texts for them.

Establishing sound texts for the plays continues to occupy the attention of a large group of scholars worldwide, with new editions appearing every few years. Where multiple varying early texts exist, choices must be made, but this can only be done if we understand the patterns of transmission that gave rise to them—their bibliography. This remains a problem because the plays were printed in a marketplace where copyright was vested in printers rather than authors, and was in any case not well developed, and no one in the chain extending from authors to book buyers was concerned with keeping records for posterity. Work on these problems goes back to the 18th century, and we now know a lot about theatres, companies, writers, printers, booksellers, and book buyers in general, but in the form of general tendencies, not hard-and-fast rules, and rarely enough about particular cases. Here we are compelled to rely on scanty documentary evidence, extended with ingenious speculation, guided by general knowledge of the practices of the day, and on close scrutiny of particular passages.

The paper will introduce a new line of evidence about the bibliography of these plays from computational stylistics. The methods are comparative and cumulative, subject to the checks against bias and the estimates of underlying reliability provided by statistical models. This provides ways of arbitrating between existing explanations, and some entirely new perspectives as well, such as systematically comparing degrees of likeness between versions.

The oldest and perhaps still dominant theory about the place of the Folio in the Shakespeare textual constellation is that the compilers of the volume, John Heminges and Henry Condell—close associates of Shakespeare, and both mentioned in his will—were able to collect fair-copy authorial manuscripts of the plays for their venture. Heminges and Condell in the preface to the Folio contend that earlier versions were nothing but ‘diuerse stolne and surreptitious copies, maimed, and deformed by frauds and stealthes of iniurious impostors’, whereas in their volume, Shakespeare’s works are ‘offer’d to your view cured, and perfect of their limbes; and all the rest, absolute in their numbers as he conceived them’.

Many scholars have followed Heminges and Condell’s lead on the illegitimacy of quarto versions of the plays, arguing that these texts derive from actors or audience members reconstructing dialogue from memory or from shorthand notes. Others have proposed that the non-Folio versions are distinct early authorial versions, pirated manuscript copies, acting versions as opposed to the reading versions printed in the Folio, or versions by other dramatists later adapted by Shakespeare for the Folio version. In this paper I focus on two of the Folio plays with earlier printed analogues, applying computational-stylistic methods to test rival theories about the nature of the Folio text and its relationship to earlier versions.

In applying computational stylistics to bibliographical questions in Shakespearean drama the study follows a chapter by Arthur F. Kinney (2009) and a book by Lene B. Petersen (2010). Kinney showed that changes between the Folio and Quarto versions of *King Lear* in uses of some alternative forms such as *which* and *that* were persistent and consistent, indicating close and purposeful revision, most likely authorial. This finding supports the most significant proposal in Shakespeare textual studies of the past 50 years, the separation of Quarto and Folio versions of *King Lear* into an early and revised state, challenging the earlier view that these were two partial witnesses to a single lost original (Taylor and Warren, 1987). Petersen aims to show patterns in some quartos that reflect memorial reconstruction, such as simplification, repetition, and transposition, but as reviews have pointed out, she fails to account for the inherent variability of her data and only rarely offers results that pass significance tests (Egan et al., 2012).

***Henry VI, Part 3***

*Henry VI, Part 3* is one of the more extreme cases of uncertainty in the Shakespeare canon. We have a prima facie likelihood that Shakespeare was at least one of the authors, given that the play is included in the Folio—titled unequivocally *Mr William Shakespeares Comedies, Histories, & Tragedies*—but scholars continue to disagree on internal, stylistic grounds about whether he wrote the whole of it, or if only part, which part, and who were his fellow authors.

As well as the Folio we have an early Octavo edition from 1595, which is different enough from the Folio version to have led some scholars to suggest that it is in fact by a different author, and should be regarded as a source for the Folio rather than an alternative version. Another group of scholars has argued that the Octavo is a memorial reconstruction of the play by actors, whereas the Folio derives from authorial ‘foul papers’ or working drafts. The latest Arden edition casts doubt on both the memorial construction and the foul papers theories, finding that neither of the two texts altogether fits its proposed category (Cox and Rasmussen, 2001, 176).

I will present findings that the 1595 Octavo is by the same author or authors as the 1623 Folio version, but the nature and extent of the differences vary greatly from act to act, with the two versions of Act IV in particular being as far apart as any two versions of an act of any of the Folio-Quarto pairs tested to date. The first two components of a Principal Component Analysis with very common words as variables and a large background set of plays in segments as observations will be used to establish a two-dimensional space as a basis for calculating distances between different versions of the same act.

***Hamlet***

Up to the early 19th century, only one Quarto of *Hamlet* was known, the one printed in 1604, but in 1823 a second separate printing was discovered, from 1603, half as long as Q2 (1604) but with a hitherto unknown scene and many alternative readings. The consensus view is that this is a memorial reconstruction of the play, but an older view that Q1 represents a much earlier authorial version of the play is revived in a recent book by Terri Bourus. She argues the case from stage history, printing history, and comparative readings of the various versions.

Computational stylistics offers an opportunity to bring some entirely different evidence to bear. Competing views put the date of composition of Q1 in the late 1580s or the late 1590s. Stylistic analysis is effective in classifying plays by date over this sort of span; the language of the drama changed quickly in this period, consistently and collectively, to the point that plays can be reliably assigned between half-decades separated by a half-decade (Craig, 2013). In later plays, words like *very* and *most* are suddenly more common, as is the indefinite article, and words like *doth*, *hath*, and *thou* forms retreat. In this paper I will show the results of dating classification tests with function-words data, lexical-words data, and word n-grams.

A late 1580s date for Q1 would undermine the idea that *Hamlet* is part of the moment of transformation in Shakespeare’s drama in 1599–1600, which is central to a number of accounts of his development and of the development of early modern culture more generally (Shapiro, 2005; Bloom, 1998; Kermode, 2000). A late 1590s dating would lend weight to the idea that an earlier lost version of the Hamlet story existed but by a different author, possibly Thomas Kyd.

**References**

**Bloom, H.** (1998). *Shakespeare and the Invention of the Human*. Riverhead Books, New York.

**Bourus, T.** (2014). *Young Shakespeare’s Young Hamlet: Print, Piracy, and Performance*. Palgrave Macmillan, New York.

**Cox, J. D. and Rasmussen, E.** (eds). (2001). *King Henry VI, Part 3*. In *The Arden Shakespeare*, 3rd series. Methuen Drama, London.

**Craig, H.** (2013). ‘The Date of *Sir Thomas More*’. *Shakespeare Survey,* **66**: 38–54.

**Egan, G., et al.** (2012). ‘VI. Shakespeare’. *The Year’s Work in English Studies,* **91**: 328–507.

**Kermode, F.** (2000). *Shakespeare’s Language*. Allen Lane, London.

**Kinney, A. F.** (2009). ‘Transforming King Lear’. In Craig, H., and Kinney, A. F. (eds), *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press, pp. 181–201.

**Petersen, L. B.** (2010). *Shakespeare’s Errant Texts: Textual Form and Linguistic Style in Shakespearean ‘Bad’ Quartos and Co-Authored Plays*. Cambridge University Press, Cambridge.

**Shakespeare, W.** (1623). *Mr William Shakespeares Comedies, Histories, & Tragedies*. London.

**Shapiro, J.** (2005). *1599: A Year in the Life of William Shakespeare*. Faber, London.

**Taylor, G. and Warren, M.** (eds). (1987). *The Division of the Kingdoms: Shakespeare’s Two Versions of ‘King Lear’*. Oxford University Press, Oxford.