SHAKESPEARE, COMPUTERS, AND THE MYSTERY OF AUTHORSHIP

Hugh Craig, Arthur F. Kinney, and their collaborators confront the main unsolved mysteries in Shakespeare's canon through computer analysis of Shakespeare's and other writers' styles. In some cases their analysis confirms the current scholarly consensus, bringing long-standing questions to something like a final resolution. In other areas the book provides more surprising conclusions: that Shakespeare wrote the 1602 Additions to *The Spanish Tragedy*, for example, and that Marlowe, along with Shakespeare, was a collaborator on *Henry VI, Parts 1 and 2*. The methods used are more wholeheartedly statistical, and computationally more intensive, than any that have yet been applied to Shakespeare studies. The book also reveals how word-patterns help create a characteristic personal style. In tackling traditional problems with the aid of the processing power of the computer harnessed through computer science, and drawing upon large amounts of data, the book is an exemplar of the new domain of digital humanities.

HUGH CRAIG is Professor of English at the University of Newcastle, Australia, where he also directs the Centre for Literary and Linguistic Computing. He has published books on Sir John Harington and Ben Jonson, and articles on English Renaissance literature and on computational stylistics.

ARTHUR F. KINNEY is Thomas W. Copeland Professor of Literary History at the University of Massachusetts at Amherst, and Director of the Massachusetts Center for Renaissance Studies. His most recent books are *Shakespeare's Webs* (2004), *Challenging Humanism: Essays in Honor of Dominic Baker-Smith* (2005), and *Shakespeare and Cognition* (2006).

SHAKESPEARE, COMPUTERS, AND THE MYSTERY OF AUTHORSHIP

HUGH CRAIG AND ARTHUR F. KINNEY

CAMBRIDGE
UNIVERSITY PRESS

Contents

vi   *Contents*

Figures

viii   *Figures*

Tables

Contributors

HUGH CRAIG is Professor of English at the University of Newcastle, Australia, where he also directs the Centre for Literary and Linguistic Computing. He has published books on Sir John Harington and Ben Jonson, and articles on English Renaissance literature and on computational stylistics.

ARTHUR F. KINNEY is Thomas W. Copeland Professor of Literary History at the University of Massachusetts at Amherst, and Director of the Massachusetts Center for Renaissance Studies. His most recent books are *Shakespeare's Webs* (2004), *Challenging Humanism: Essays in Honor of Dominic Baker-Smith* (2005), and *Shakespeare and Cognition* (2006).

PHILIP PALMER is a doctoral student at the University of Massachusetts at Amherst, and Curator of the Massachusetts Center for Renaissance Studies. His research interests include representations of remote space in Renaissance drama and poetry, travel literature, pamphlet culture, and manuscript studies.

TIMOTHY IRISH WATT is a doctoral candidate at the University of Massachusetts at Amherst, where he has been the Walter T. Chmielewski Fellow for *English Literary Renaissance*.

Preface and acknowledgments

This project began in 2001 when one of the editors, Arthur Kinney, spent a period as a research visitor at the University of Newcastle in New South Wales, where the other editor, Hugh Craig, works. A collaborative project was hatched. It was to combine Kinney's knowledge of Shakespeare with Craig's familiarity with numbers, and thus pioneer a Shakespearean computational stylistics. Kinney would supply the questions and Craig would furnish numerical results. In what they jointly wrote they would keep in mind an audience that had no interest in arcane statistics or in interminable tables of figures. Authorship would be the core of the enterprise. Coworkers were recruited from Kinney's Massachusetts Center for Renaissance Studies in Amherst, as was a programmer and research assistant from Craig's Centre for Literary and Linguistic Computing in Newcastle, so it has been an American–Australian and a Center–Centre collaboration.

Computational stylistics stems from the work of John Burrows, beginning in the 1980s. He was convinced that the smallest elements of literary language (down to very common grammatical words such as *and* and *but*) had things of stylistic interest buried in them, and he sought to bring latent patterns in their use to light, through multivariate statistical procedures analysing the word-count information from large tracts of machine-readable text. The new method revealed a patterning that pervaded all levels of language and could be measured. Readers, however perceptive they may be, have only one lifetime to read in, and remember what they read selectively and imperfectly. The computer, on the other hand, is brutally simple in its relationship to the text, but it has superhuman powers of memory, and can deal with what it knows with unthinkable rapidity (and always in a predictable and repeatable way). With its help, Burrows' computational stylistics offers new views of literary landscapes. In its most obvious application, where experts disagree on the authorship of a text and there is no external evidence to help, computational stylistics can offer an

objective arbitration. It can also work in a more exploratory way, by looking for unexpected patterns and quirks in the dataset itself, rather than testing a hypothesis. In the book we have examples of both approaches.

One might expect a stubborn contradiction between language, and other artistic means, and computation. One of the purposes of the book is to show that there is in fact a considerable sympathy between them. Early Modern plays were written to be seen or read, not subjected to tabulation and statistical analysis, but the plays do work by frequency and distribution as well as by the impact of the individual word, speech, and scene. This means that computational work (we believe) opens a new gateway to complexity and nuance in language, rather than running roughshod over them.

Beyond the immediate questions of Shakespeare authorship the work in the book is meant as a contribution to the larger question of stylistic individuality. Computational stylistics offers abundant evidence that writers leave subtle and persistent traces of a distinctive style through all levels of their syntax and lexis. This brings to the fore a central paradox of language. Speakers and writers share the words they use in a given language. They could not communicate otherwise. Yet from that common set speakers and writers make individual selections that persist across all their uses of the language. They create a personal and identifiable style from within the common language. Computational analysis reveals the richness of this variation within the dialogue of Shakespeare and his contemporaries. It persists even when dramatists strive to create their own fictional linguistic individualities in characters. Hal, Falstaff, and Hotspur do have their own languages, but underlying them all is a Shakespearean idiom, which means they are all distinct from Jonson, Marlowe, or Middleton characters. This idiom is so powerful and persistent that even the computer can detect it. We can even turn this around and say that it wasn't until the computer came along that we could properly appreciate some aspects of this miraculous secret working of language in these very familiar plays and characters.

We believe the book is more wholeheartedly statistical, and computationally more intensive, than any previous Shakespeare study. It embraces statistical principles such as testing samples with known classifications before trying the measures on doubtful cases. Where possible we have used multiple separate tests for a given problem. We hope we have incorporated a fully fledged, built-in scepticism about the reliability of the results. Extensive work has already been done in the area of quantitative study of Shakespearean authorship, by scholars from the literary world like

MacDonald P. Jackson, Gary Taylor, and Brian Vickers, and by researchers from the computer-science and statistics side like Ward E. Y. Elliott, Thomas Merriam, M. W. A. Smith, and Robert J. Valenza. We have tried to unite the literary and statistical sides through our collaboration, and to build on previous work to give the most holistic computational modelling of style so far. We aim to resolve a number of questions in the Shakespeare canon, so that the business of interpretation, which is so often stymied by uncertainty of authorship, can proceed. Vickers' book *Shakespeare, Co-Author* (2002) has already done this for the division of five collaborative plays among their various authors; we hope to bring a similar level of confidence to the question of the Shakespearean authorship of *Edward III*; *Arden of Faversham*; the Additions to *The Spanish Tragedy*; the Hand-D Addition to *Sir Thomas More*; *Edmond Ironside*; the Folio *King Lear*; and *Henry VI*, *Parts 1 and 2*.

Readers will want to know how many texts we have included, to judge the basis on which we generalize about authors and trends, and will be curious about the nature of the texts that underlie the whole enterprise. We use early printed versions as copy-texts, to minimize the effects of modern editing and to open up the corpus to two or more editions where these differ significantly. Each text is tied therefore to a single early witness. Consequently spellings are Early Modern and highly variable. We standardize selected function words to modern usage. For the rest, a process has been developed within the software we use for word-counting to group variant spellings as teams, collecting the different forms of the same word under a single head word, which can then form the basis for counting. Thus instances of 'folly', 'follie', and 'folie' are all counted under the head word *folly*. The corpus we have assembled for the book consists of 165 Early Modern English plays, around 3.25 million words of dialogue in all (they are listed in Appendix A of this book). Of these, 138 are from a more narrowly defined 'Shakespearean' period, which one might define as 1580–1619 (the four decades in which Shakespeare is presumed to have been active). Some of these plays are of mixed or disputed authorship, and are the subject of investigation, so cannot form a set of standards for the core purpose of the study, the defining of Shakespearean authorship. For this we need single-author, well-attributed plays to serve as exemplars of Shakespeare's style and those of his contemporaries. There are 112 such plays within the 138 in our corpus. The *Annals of English Drama* (1964 edition) lists 174 surviving well-attributed single-author plays from this period. The corpus thus contains 112 out of 174 – just under two-thirds – of all the available usable plays for attribution purposes. It includes

complete sets of the surviving Shakespeare, Marlowe, Jonson, Middleton, and Webster plays: complete, at least, according to a conservative standard of what is 'well-attributed' for each writer. We have four or more plays by seven other playwrights: Lyly, Greene, Peele, Dekker, Heywood, Ford, and Fletcher; and three each by Robert Wilson, Chapman, and Shirley.

As to methods and procedures, our aim in the chapters that follow has been to explain the steps we took to get from these texts to the results with enough detail so that anyone wishing to replicate the findings is able to do so. The patterns we have uncovered should be robust enough to survive the variations that will arise from using different texts and software, provided the same basic procedures are followed. (The question of whether we have made the right judgments in choosing among the possible procedures, and in setting the various parameters, is quite another matter.)

Warren Stevenson's *Shakespeare's Additions to Thomas Kyd's* The Spanish Tragedy (Lewiston, NY: Edwin Mellen, 2008) came to our attention after this book was in production. We regret not being able to make use of Stevenson's amplified case for Shakespeare's authorship of the Additions here. We quote extensively from Stevenson's 1968 article on the topic, which presents a briefer version of his evidence, below.

The project was made possible by funding from the National Endowment for the Humanities and the Australian Research Council, for which we are grateful. Sarah Stanton at Cambridge University Press has done everything one could ask of a commissioning editor, asking searching questions of us at a formative stage and then championing the project when it ran into controversy. We have also had the assistance of a great many expert and generous co-workers, and we would like to thank all of them. Graham Christian, Youngjin Chung, Kimberly C. Elliott, Kevin Petersen, and Anne-Marie Strohman were all involved in the early stages of the project, and have contributed to it in numerous ways. R. Whipp developed the software for the Intelligent Archive, which has been the 'two-handed engine' making light work of the word-counting lying behind the quantitative studies. Alexis Antonia and Ruth Lunney have prepared play texts for the project. Penny de Sylva of Oxford University Press arranged for us to use the electronic *Oxford English Dictionary* as a source for the variant spellings incorporated in the Intelligent Archive. John Burrows, Andrew Craig, David Craig, Henry Craig, Mark Gauntlett, David Hoover, John Jowett, and Ruth Lunney have all read chapters and saved us from many errors. We would also like to acknowledge the generosity of Mac Jackson, Tom Merriam, and Brian Vickers in letting us see unpublished results and sending hard-to-find publications in authorial attribution. We also owe

a debt to the three anonymous readers for Cambridge University Press, who gave us a greatly improved sense of what an audience for a book like this might expect and much detailed guidance besides. It only remains to say that the errors that remain are entirely the responsibility of the four authors of this volume.