The Marriage of Philology and Informatics

As part of the British Academy Literature Week in October 2009, **Sir Brian Vickers FBA** *discussed 'The Authors of King Edward III', in conversation with Professor Laurie Macguire. Here he describes how information technology can be harnessed to the literary study of authorship attribution.*

IN THE LATE 5th century AD, Martianus Capella wrote a book in prose and verse called *De Nuptiis Philologiae et Mercurii*, an allegory of the seven liberal arts, which had a considerable influence on medieval culture. It begins with two books describing the ascent to heaven, apotheosis, and marriage of Philology and Mercury, based on Neoplatonic doctrines on the ascent of the soul. Then follow seven books describing each subject by an elaborately described female personification, three books being given to Philology's word-based arts, *Grammatica*, *Dialectica*, *Rhetorica*, and four to Mercury, dealing with mathematical concepts: *Geometria*, *Arithmetica*, *Astronomia*, and *Harmonia*. For centuries in the West, education began with Philology's *trivium* and was completed by the *quadrivium* of Mercury, the whole forming a complete curriculum.

The word *philology* briefly entered the English language in about 1386 (thanks to a jocular reference in Chaucer's *Merchant's Tale*), and when it re-appeared in the 17th century it meant the love of learning and literature. It was not until the 18th century that *philology* established itself to describe 'the study of the structure and development of language', as the *Oxford English Dictionary* defines it. The same authority dates the introduction of *informatics* to 1967 (from the Russian *informátika*, 1966), and cites a definition of it as 'the discipline of science which investigates the structure and properties (not specific content) of scientific information', also known as 'information science'.

Authorship attribution

My research interests in recent years have been in the discipline of authorship attribution study, traditionally the domain of philology. Many of its practitioners have used their remarkable linguistic knowledge to reject spurious attributions, as did Varro and Aulus Gellius with the canon of Plautus, Valla with the 'Donation of Constantine', Erasmus with works attributed falsely to Jerome, Isaac Casaubon with the supposedly Mosaic corpus of Hermes Trismegistus, or Richard Bentley with the Epistles of Phalaris. These scholars were able to expose forgeries and falsely-attributed works by drawing on historical knowledge and a sensitivity to language and style, attributes that are still essential in authorship studies.[1] But when the problems involve not texts written centuries after their supposed date but anonymously-published or pseudonymous works written in the same epoch and in the same genre, appropriate analytical techniques must be devised.

In the 19th century several Shakespeare scholars, wanting to identify his share of the two plays he co-authored with John Fletcher (*King Henry VIII*, *The Two Noble Kinsmen*), developed some simple quantitative methods for analysing verse style, counting the proportion of lines having more than ten syllables, or having 'light' or 'weak endings'. Although the mathematics was elementary, the method successfully distinguished the two authors, and their assignment of individual shares is still accepted.[2] In the 20th century another problem in identifying co-authors' shares was solved with a far more advanced Mercurial discipline, statistics. In 1787–8 three New York newspapers carried *The Federalist*, a series of articles advocating the ratification of the United States Constitution. The initiator and main author was Alexander Hamilton, who in 1789 became the first Secretary of the Treasury, assisted by James Madison, who in 1808 became the fourth President. Subsequent collected editions of the *Federalist* papers identified the authors of individual essays from notes left by Hamilton and Madison, but their lists diverged, both claiming authorship of a dozen essays. Two American statisticians, Frederick Mosteller and David L. Wallace, chose as authorship markers a set of 90 'grammatical' or 'function' words (such as prepositions, conjunctions, definite and indefinite articles), and 60 'content' words. They manually computed the rates of relative occurrence in all 146 essays, and subjected the results to a Bayesian statistical analysis which computed the relative odds of assigning authorship.[3] The majority of the disputed papers were ascribed to Hamilton.

Elizabethan and Jacobean drama

Given machine-readable texts of *The Federalist*, the labours of Mosteller and Wallace in compiling this list of word occurrences could be performed today in a few seconds, and several statistics packages could instantly calculate probabilities. But the problem they faced was relatively simple, in that only two authors were involved, and they had adequately long text samples of each (also, Madison seldom used the preposition 'upon'). In the field in which I mostly work, drama written and performed in London between 1580 and 1642, issues are more complex. First, the high demand for new plays among the competing theatre companies meant that in about half the recorded instances two or more authors collaborated in supplying playbooks, which then belonged to the companies. Secondly, when the companies sold plays to a printer, either to raise cash or publicise an imminent revival, the title-pages advertised themselves rather than the authors. Marlowe was not credited with *Tamburlaine the Great*, nor Shakespeare and Peele with *Titus Andronicus*. Thirdly, the plays' language was restricted by the audience's ability to understand a play at first hearing, and by the dramatists' use of unrhymed decasyllabic verse. The unfettered medium of prose allowed a Nashe or a Burton to develop highly individual styles, but the more limited scope of the iambic pentameter, and the nature of the audience, were stylistic levellers. Finally,

since dramatists worked intensely in such a small theatrical world, some of them acting in their own or other people's plays, mutual linguistic influence was unavoidable. Identifying the perhaps two or three authors of an anonymously published play is a far harder problem than deciding which of two known authors wrote a *Federalist* essay.

In authorship studies, as in all rational enquiries, it is vital to design a properly structured method, based on a sound theory, and using procedures that can be replicated by other researchers. The most important 'philological' decision is to identify those aspects of language that will best reveal individuality. Given the speed with which computers can search, count, and sort machine-readable texts, quantities of data are quickly supplied. But it is no use moving from language to the computational level, producing columns of numbers, if you cannot show any organic connection with the words of the text. Elizabethan drama is a genre in which authors are not immediately visible: they speak through their characters, who are individualised according to gender, age, social class, and dramatic function. A simple computation of function words, however elaborately sifted by statistical procedures, may tell you something about the characters but cannot reliably indicate authorship.[4] Searching by content words suffers from the liability that plays dramatising the prolonged battles of medieval English history, say, will share a vocabulary of swords, armour, horses, combat, blood, wounds, death, victory. It may be fruitful to use atomistic approaches, based on single words, in studying non-fictional works, or even prose fiction written by single authors, but for drama we need a method which will respect what Saussure described as the linearity of language, the distinctive feature by which we add word to word in order to form our utterances.

Corpus linguistics

To find a method that respects the sequential or 'joined up' nature of language, I have turned to a field in which the marriage of Philology and Informatics has been extremely fruitful, the new discipline of corpus linguistics. In 1964 the *Brown University Corpus of Present Day American English* appeared, the first machine-readable corpus, consisting of texts published in 1961 and amounting to a million words. In the four decades since then, dozens of corpora have appeared, greater in scale (the 'Bank of English', produced by Birmingham University in association with the dictionary publisher, Collins, had grown to 524 million words by 2004), and wider in scope (now covering all varieties of present-day English, and gradually reaching back into the past). The provision of these vast databases has stimulated the empirical study of language use, a field long neglected during the ascendance of linguistic theory. In addition to providing a new method for studying language as a system, corpus linguistics, being based on actual language use, can indicate the probability that certain structural patterns will appear. With automatic concordancing and other data-mining tools, corpus linguistics can reveal facts about language use never previously suspected.[5]

For authorship studies the major issue is its confirmation of a linguistic phenomenon noted by J.R. Firth in the 1950s, that we tend to use words in groups, not singly. Where Chomskyan theory holds that

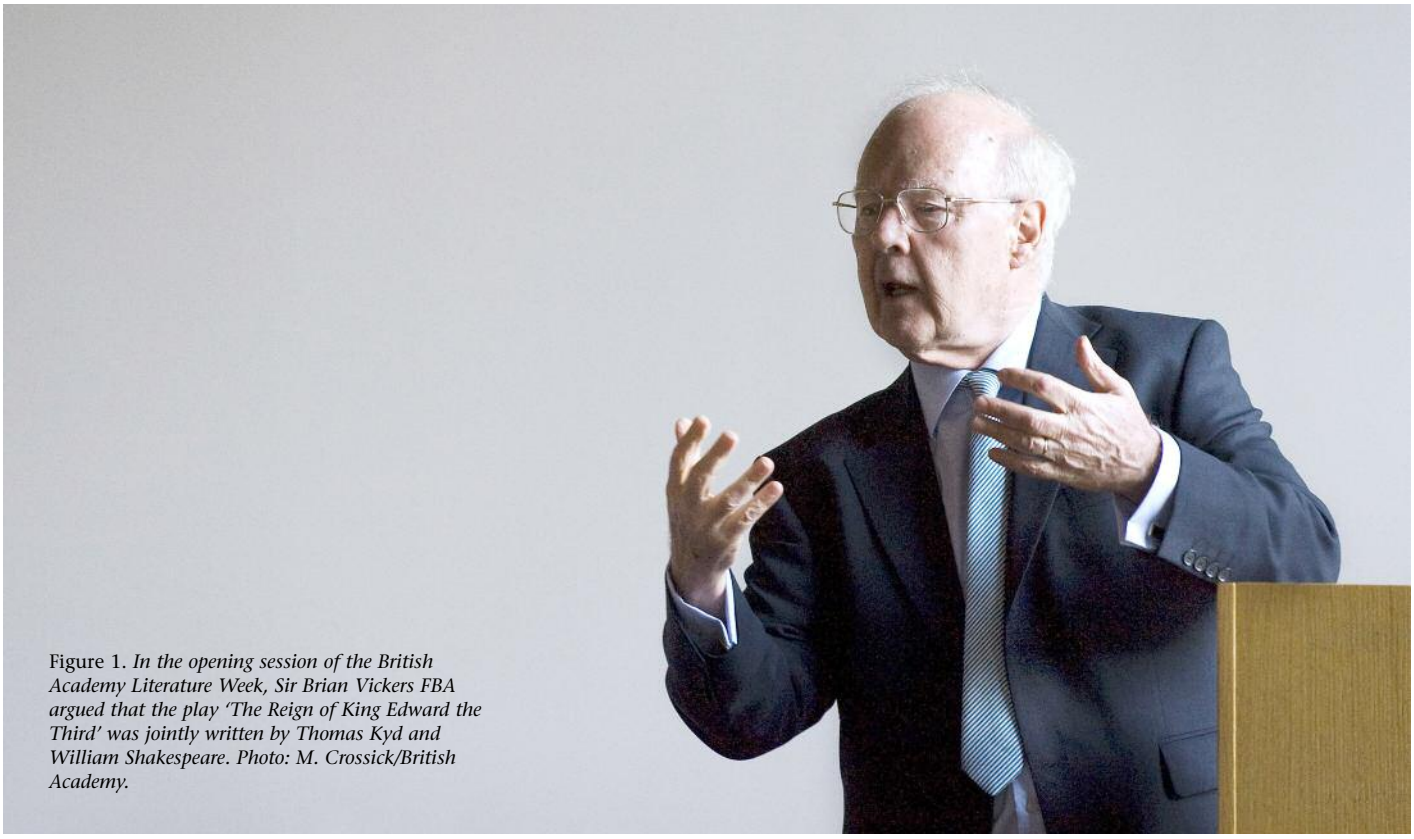


Figure 1. *In the opening session of the British Academy Literature Week, Sir Brian Vickers FBA argued that the play 'The Reign of King Edward the Third' was jointly written by Thomas Kyd and William Shakespeare. Photo: M. Crossick/British Academy.*

sentences are generated purely syntactically, based on grammatical rules and individual lexical items, corpus linguistics shows that in many cases sentences come partially lexicalised. That is, the human brain may process language one word at a time, but it also deals with word-strings, ready-made phrases or collocations in which some words frequently recur in regular combinations. The late John Sinclair, an admirer of Firth and a major force in the creation and exploitation of linguistic corpora, acknowledged the 'open-choice principle', by which the brain generates unique sentences, but defined a complementary 'idiom principle', by which 'a language user has available … a large number of semi-constructed phrases that constitute single choices, even though they might appear to be analyzable into segments'.[6] Other researchers using corpora have shown that 'formulaicity [is] all-pervasive in language data'.[7] Strangely enough, it needed modern informatics to show philology how language actually works. We are surprised to learn that in any lexicon the set phrase (or 'phraseme') is the numerically predominant lexical unit, outnumbering single words roughly ten to one, and that in one corpus about 70 per cent of its half a million words of running text are part of recurrent word combinations.

The most famous instance of such phrasal repetitions in world literature are the so-called epic formulae in Homer, 'swift-footed Achilles', 'the wine-dark sea', and so on, which are given central importance in a remarkable new database, 'The Chicago Homer'.[8] Edited by Ahavia Kahane (Royal Holloway) and Martin Mueller (Northwestern University), this multilingual database uses the search and display capabilities of electronic texts to make the distinctive features of Early Greek epic accessible to readers with and without Greek. Significantly, the editors quote John Sinclair's definition of the 'idiom principle', and argue that 'the Homeric Question is a question about phrasal repetition'. Accordingly, they have compiled a database consisting of the *Iliad* and *Odyssey*, the poems of Hesiod and the Homeric Hymns, and subjected it to skilled pre-processing. Users can easily pick out any sequence of two or more words that is repeated at least once in the corpus of Early Greek epic, 'whether the repeated sequence is a mere bit of syntactic glue such as "but he", a noun-epithet formula such as "rosy-fingered Dawn", or an extended narrative stretch, such as the story of the shroud Penelope wove for Laertes.' An extraordinary amount of sharply focused knowledge and experience has gone into creating this resource, which could well revolutionise Homeric studies. From the viewpoint of authorship attribution the most interesting findings are that 'phrasal repetition is a very pervasive phenomenon', no less than 36,000 lemma strings being repeated in this corpus, ranging in length from 2 to 123 words and in frequency from 2 to 3,152 repetitions. A search for repeated phrases longer than 40 words will produce a list with 31 hits. These figures are remarkable in themselves, and carry considerable significance for literary analysis, as one of the editors has recently shown.[9] But their cultural implications are far greater, for the marked up text of this database allows users to follow the links of repeated passages as a way of navigating 'the neural networks of bardic memory', thus simulating the competence of the ancient listener.

Repeated phrases in Elizabethan plays

Phrasal repetition may play a greater role in Homeric epic than elsewhere, but it occurs in everyday language and in written texts, including Elizabethan drama. So we need a tool that can automatically analyse the continuous text of these plays and pick out repeated collocations. We then need to check each collocation against the corpus of plays performed up to a given point, to establish whether dramatists had favourite phrasemes, and if so, whether they occur in sufficient quantities in anonymously published or co-authored texts for us to be able to identify the hands, brains, or 'neural networks' which composed them. The whole process should be automatic and replicable, to evade the longstanding objection against authorship attributions based on 'parallel passages' that the parallels were only visible to the scholar who claimed them, and could not be tested against the work of other dramatists.

An effective way of measuring an Elizabethan author's tendency to repeat his favourite phrasemes is to use one of the tools that now exist to detect students' cribbing, such as 'Pl@giarism'.[10] This program compares two machine-readable texts in parallel and automatically highlights every instance where they share identical three-word sequences. If we compile a database of all the plays staged in a dramatist's lifetime, by using a data-mining program, such as 'InfoRapid Search & Replace',[11] we can ascertain in a fraction of a second whether a collocation is widely shared or rare. Having established an author's regular stock of collocations, as found in authenticated plays, we can then turn to the anonymously-published or co-authored plays in which his hand has been suspected, and check how many times his favourite collocations occur there.

Thomas Kyd

With the help of Dr Marcus Dahl, who devised this method, I have been studying the canon of Thomas Kyd, which currently consists of only three plays, *The Spanish Tragedy* (*c*. 1587), *Solyman and Perseda* (*c*. 1590), and *Cornelia* (*c*. 1594), together with a prose work, *The Householders Philosophie* (1588). Having compared the canonical works with each other, two at a time, I was able to establish Kyd's frequent self-repetition (his Spanish and Turkish tragedies share over 80 collocations of three words or more). I then compared each of those works with three plays I had good reason to think were by him, with striking results. The bourgeois tragedy *Arden of Faversham* (*c*. 1592) has 95 unique matches with the canonical works – matches not found anywhere else in the 64 plays performed on the London stage before 1596. *The True Chronicle of King Leir and his three daughters* (*c*. 1590), shares no less than 125 unique matches with the accepted Kyd canon. Ben Jonson referred to 'sporting Kyd', suggesting that he had written comedies, and I can now attribute to him the badly-abridged romance (half the length of a normal play), *Fair Em, the Millers daughter of Manchester: with the love of William the Conqueror* (*c*. 1590), which shares 42 unique matches with the four accepted works. The number of unique matches in each case far exceeds coincidence, imitation, or plagiarism. Moreover, I can support these findings by traditional literary analysis, all three plays

sharing many elements of narrative technique and dramatic structure with the accepted canon.[12]

Whereas the use of parallel passages as an authorship test has, in the past, been criticised as subjective and unreliable, these Kyd collocations have been picked out automatically, with no prior input from me. I have had to manually check each against the corpus of 64 plays produced in the London theatres between 1580 and 1596, a tedious process but with the advantage that – as the editors of the Chicago Homer have also noted – the reader's eye can detect discontinuous collocations, larger patterns invisible to the machine. With practice anyone would achieve the same results, so that this method can be granted scientific status, being both objective and replicable. I have recently been applying it to two plays whose presence in the Shakespeare canon has long been debated; my work in progress suggests that the Kyd canon will soon be enlarged further. The proper marriage of philology and informatics has the power to solve many problems yet.

Sir Brian Vickers is Distinguished Senior Fellow at the School of Advanced Study, University of London, and a Fellow of the British Academy.

An audio recording of the discussion is available as a podcast from www.britac.ac.uk/medialibrary/

British Academy Literature Week

Leading academics, writers and practitioners came together in a series of linked events which made up the British Academy's highly successful Literature Week. Audio recordings of all the events are available as podcasts from www.britac.ac.uk/medialibrary/

Monday 19 October:	(**A**) *Sir Brian Vickers FBA preparing with Professor Laurie Maguire for their conversation on* 'The Authors of King Edward III'*;* (**B**) *Dr Tom Lockwood delivering the Chatterton Lecture on* 'Donne, by hand'*.*
Tuesday 20 October:	(**C**) *Sir Christopher Ricks FBA in conversation with Professor Hermione Lee FBA on T.S. Eliot ('and others');* (**D**) *Professor Marina Warner FBA introducing* (**E**) *Professor Robert Crawford, who gave the Warton Lecture on* 'T.S. Eliot's daughter'*.*
Wednesday 21 October:	(**F**) *Michael Lesslie, Professor Jonathan Bate FBA and Lindsay Posner discuss* 'American writers on the English stage'*;* (**G**) *Professor Christopher Bigsby delivering the Sarah Tryphena Phillips Lecture on* 'Arthur Miller: poet of the theatre'*.*
Thursday 22 October:	(**H**) *Josephine Hart, Kenneth Cranham, Elizabeth McGovern and Charles Dance perform in the Josephine Hart Poetry Hour;* (**I**) *Professor Frank McGuiness discussing poetic theatre* – 'Beyond verse?'

Notes

1 See Harold Love, *Attributing Authorship: An Introduction* (Cambridge, 2002).

2 See Brian Vickers, *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays* (Oxford, 2002).

3 See F. Mosteller & D.L. Wallace, *Inference and Disputed Authorship: The Federalist* (Ithaca, NY, 1964; Stanford, CA, 2008).

4 The sensitive study by John Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford, 1987), differentiates characters on the basis of the function words they use.

5 See the excellent survey by Graeme Kennedy, *An Introduction to Corpus Linguistics* (London 1998).

6 See John Sinclair, *Corpus, Concordance, Collocation* (Oxford, 1991), p. 109.

7 Alison Wray, *Formulaic Language and the Lexicon* (Cambridge, 2002), p. 13.

8 See www.library.northwestern.edu/homer/understandinghomepage.html

9 See Martin Mueller, *The Iliad*, second edition (London, 2009), especially chapter 6.

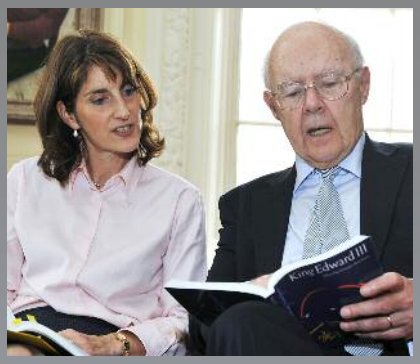
10 See www.personeel.unimaas.nl/georges.span/Plagiarism

11 See www.inforapid.de

12 For a preliminary account of these researches, see Brian Vickers, 'Thomas Kyd, Secret Sharer', *Times Literary Supplement* (18 April 2008), pp. 13–15; and for an independent validation of the results, see <data.at.northwestern.edu.mht> postings of 18 and 23 August 2009.

BRITISH ACADEMY LITERATURE WEEK, OCTOBER 2009

A

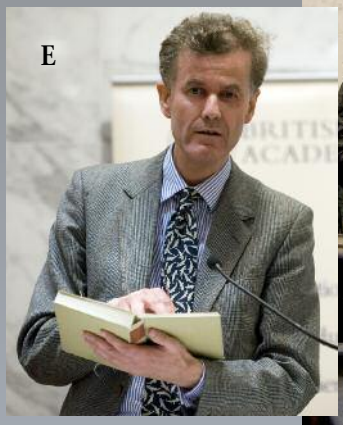


B

C

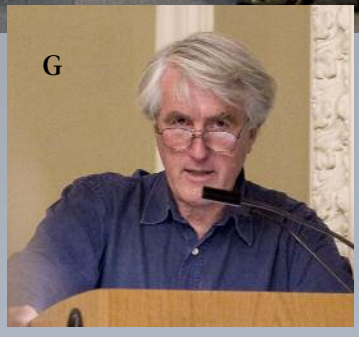
D

E



F

G



H

I

Photos: M. Crossick/British Academy