

nance a far more upsetting one: "With her, on her - what you will". We all choose the pronoun that we will (or want), but the imagined scene created by "on" causes Othello to collapse into madness. This may be the most devastating use of a preposition in world literature, but it proves that function words are chosen. Because function words combine with content words to create meaning, separating our either class destroys the possibility of understanding the meaning of language or literature.

For many readers this is the fundamental weakness of quantitative attribution study, that it abandons meaning. The 100 or 500 words used most frequently by the target author(s) are reduced to mere items to be counted, without consideration of context or meaning. But many words have many meanings, which can only be construed by the context of use, and poets create unique contexts. Romeo's first words on seeing Juliet - "O she doth teach the torches to burn bright" - yields one word for the dataset, *torches*. Cleopatra's words on Antony's death - "The soldier's pole is fallen" - yields two, *soldier* and *pole*. In her mouth that is not just any pole but a metaphor for their intimacy. Abandoning the sequential interplay between words and treating them as separate items for computation destroys the possibility of meaning, and of those extra meanings that are the life of drama. Of course, given the exponential increases in computing power and data storage over the past sixty years, it may well be that computerized textual analysis will ultimately go beyond treating word frequencies in isolation and will be able to cope with the full range of language. But that time has not come, as the *Authorship Companion* shows only too well.

Here I present our findings for two methods, one elaborate, the other suspiciously simple. The first, attributing parts of *Henry VI* to Marlowe, was originally published in a journal of signal processing. A revised version, by Santiago Segarra, Gabriel Egan, Mark Eisen and Alejandro Ribeiro, appeared in *Shakespeare Quarterly* (2016) and was instantly endorsed by Gary Taylor, although it had received no independent review. The new method, "Word Adjacency Networks", is based on the dubious assumption that dramatists can be identified solely by their function words. The authors believed that "the choice, the frequency, and (crucially for our method) the relative placing of function words ... appear to be an unconscious set of preferences specific to an author" (that fallacy yet again). The term "placing" means the number of content words that intervene between each preposition or pronoun, say. The opening sentence of *Paradise Lost* would be reduced to "Of XXX and the X of that XX, whose XXXX into the X" and so on, the intervals between those words being mathematically significant. I thought this a fantastic claim when I first read it, and sustained scrutiny of all the publications by this group has only confirmed that impression.

The four authors illustrated their method by comparing two brief passages by Shakespeare and Dekker, highlighting the function words that they happen to share: *with, one, and, in*. They arranged these four words into two square-shaped "word adjacency networks" with one word at each corner - or rather, node - and with an arrow-headed line - or edge - linking them in various repetitions. There were more of these in their chosen Shakespeare excerpt (Claudius's self-incriminating attempt to justify his over-hasty marriage) than in the Dekker, and for each repetition they "raised the existing edge's weight by one, taking it from 1 to 2", and so on. According to their computations "Shakespeare would be most likely to use *in* shortly after he used *with* ... for the preference is encoded by the weight of 3 (the highest in this WAN)", so providing a reliable "expectation of Shakespeare's word-choices in an unseen piece of his writing".

As a non-mathematician I was disturbed by their treating language as if it were an electrical circuit and by their belief that the random sequence of function words in Claudius's hypocritical oratory could tell us anything about Shakespeare's "word-choices". I call

it random because it is evident that function words complement content words but are subordinate to them. First we think of the things or people, then we put them in relation to each other. Pervez Rizvi checked his database, which showed that a character's choice of a function word is often determined by a lexical word he has just used. He then evaluated the mathematics of Segarra's method, finding that it boiled down to half a dozen formulae: "these are so unfit for the use made of them that one formula has to pretend that some words are present in a text when they are absent, while another has to pretend that they are absent when they are in fact present".

The essay duly divided *3 Henry VI* between Shakespeare and Marlowe, awarding nine scenes to Marlowe and thirteen to Shakespeare, with six scenes "too close to call". A recent study by other members of the *New Oxford Shakespeare* attribution team, using another unsound method ("Zeta"), had produced a different result, attributing sixteen scenes to Marlowe and twelve to Shakespeare. Our four authors congratulated themselves on having corroborated the "measurable amounts of Marlowe's writing" in this play, except for "scenes 1.2, 2.4, 4.3, 4.4, 4.6, and 5.7". Although blind to the massive discrepancy, they did raise a problem that is likely to occur more often: how can we adjudicate attributions when two quantitative methods give different results? Must we decide a literary issue according to the rival claimants' use of maths, or statistics?

One thing we can always do is study the text, looking for visible properties that might differentiate one dramatist from another, such as their prosody. The first generation of dramatists for the public theatre - Marlowe and the other "university wits" - used the fixed ten-syllable line with little deviation. In the late 1580s Kyd pioneered the adding of an extra syllable, the so-called feminine ending (a term taken over from French metrics), as in "To be or not to be, that is the quest'ion". Shakespeare used this expressive device increasingly through his career, together with more flexibly verse movement. In 1850 James Spedding used both features to demonstrate that *Henry VIII* had been co-authored by Shakespeare and Fletcher. His allocation of the scenes to each dramatist is accepted to this day. The proportion of feminine endings in Marlowe runs from 1.2 to 3 per cent, while that of Shakespeare in *3 Henry VI* is 10.7 per cent, way beyond Marlowe's metabolism. Anyone with a reasonable ear for verse can tell the two dramatists apart.

A second well established method is to compare dramatists' phraseology (phrases of three or four consecutive words). While single words are insufficient to identify a dramatist, longer word strings can - indeed, corpus linguistics has shown that all language users "chunk" words into short sequences. I used the Rizvi marked-up corpus to identify the unique phrases matching *3 Henry VI*, that is, phrases which occur in it and in only one other play. I checked the first thousand and recorded 115 matches with Shakespeare, all of high quality. There were only twenty-five commonplace matches with Marlowe. (I discounted fourteen with *Edward II* since it postdates Shakespeare's play.) This is doubtless a rough-and-ready measure but all other evidence suggests that *3 Henry VI* is a play of sole authorship.

Although he endorsed Word Adjacency Networks, Gary Taylor preferred a simpler approach. Middleton's increased share of *Macbeth* in the recent edition derives from a method that he had invented himself, called "micro-attribution". Where other scholars use segments of 2,000 or 5,000 words, Taylor claimed he could determine the authorship of a speech by Hecate in four rhyming couplets, or only "sixty-three consecutive words". To do so one must "analyze every word and every possible combination of words", starting with two consecutive words, known as a bigram. (Taylor claimed that this sixty-three word speech contains 124 bigrams, but the correct figure is sixty-two.) In Hamlet's soliloquy, then, that would give "To be", "be or", "or not", "not to", and so on. Then we would move on to trigrams, that is

“
I was
disturbed by
their treating
language as if
it were an
electrical
circuit

"To be or", "be or not", "or not to", "not to be", and so on, then to strings of four words, then five. Taylor claimed to have obtained "905 individual data points" from only sixty-three words, a new miracle of the loaves and fishes. On first view I thought this a daft method, treating words like counters in a board game and creating meaningless word-units, which the player would search for in texts by other authors. Taylor solemnly applied it to passages of matching length and verse form in plays by Middleton and Shakespeare, and by a lengthy process of calculation involving very small matches (nine to eight, or six to four), he assigned Hecate's speech to Middleton. No reputable scholar would accept attributions made on such Lilliputian samples.

Pervez Rizvi criticized several aspects of this theory. First, it was naive of Taylor to think that his various combinations enlarged his database. "A truly large sample consists of many independent observations. Observations created by combining other observations are not independent". Independent observations "can only be taken from the source - in this case the text of *Macbeth*". Secondly, Taylor couldn't possibly have analysed every verbal combination, for the sixty-three-word *Macbeth* passage would generate tens of millions of matching collocations with other plays. But Taylor identified a mere eighteen matches, without explaining in what way they were representative (thirteen of them are not unique to Shakespeare and Middleton). His list includes such banal phrases as "so - Ay, sir" and "Ay, sir, all", but Taylor claimed that his results proved Middleton's authorship. Wishing to make an independent check of the two dramatists' relative positions, Rizvi reduced Taylor's sixty-three-word window to ten words and got 759 results. When these matches were ranked in frequency by dramatist, Shakespeare topped the list, with Middleton in eighth place. Rizvi judged that "his faux-scientific method would be laughed out of the room at a scientific gathering". Sadly, Taylor and his pupils have persuaded major scholarly journals to publish essays using it, and in the *Authorship Companion* it enables the transfer of the "Fly scene" in *Titus Andronicus* and parts of *All's Well to Middleton*. Seldom can major issues have been decided by such a questionable method.

Over and above its claimed attributions, the seemingly professional *Authorship Companion* contains many defects. David Auerbach described it as a "shambolic" collection, an "unreliable grab bag" in which contributors failed explicitly to declare their criteria, borrowed unsuitable methods from other disciplines (machine learning, biochemistry) and violated key principles of statistics. He was particularly scathing about "the poverty of the input data. By restricting their analyses to a handful of primitive signals such as word frequency and word succession, many of these researchers end up coating fundamentally simple (and untenable) findings in a statistical glaze". This *Authorship Companion* is unfortunate proof that scholars, journal editors and publishers in the Humanities are prone to being abused by pseudo-scientific methods.

Returning to the *New Oxford Shakespeare* map of the canon, those encroaching colours will be permanent stains on the edition, for every attribution is false. Oxford University Press has a proud record as the world's leading publisher of scholarly editions of English literature. The trust that senior editors placed in Gary Taylor has been repaid with an opportunistic bundle of untested methods set loose on the greatest author in our language. Shakespeare is not just a national, but an international treasure and it is tragic to contemplate the damage done to culture in general by these editions being used to teach students, and being sold in bookshops to unsuspecting laymen. The Press has just commissioned the *New Oxford Marlowe*. Among its editors are members of Taylor's editorial team, and rumour suggests that it will include the *Henry VI* plays. Many people will fervently hope that on reflection the editors will think it enough to have ruined one major author's canon. ■