CULTURE

The Ghost's Vocabulary

How the computer listens for Shakespeare's "voiceprint"

By Edward Dolnick

OCTOBER 1991 ISSUE

In 1842 literature and science met with a thud. Alfred Tennyson had just published his poem "The Vision of Sin." Among the appreciative letters he received was one from Charles Babbage, the mathematician and inventor who is known today as the father of the computer. Babbage wrote to suggest a

correction to Tennyson's "otherwise beautiful" poem—in particular to the lines "Every moment dies a man, /Every moment one is born."

"It must be manifest," Babbage pointed out, "that, were this true, the population of the world would be at a standstill." Since the population was in fact growing slightly, Babbage continued, "I would suggest that in the next edition of your poem you have it read: 'Every moment dies a man,/Every moment 1-1/16 is born.'" Even this was not strictly correct, Babbage conceded, "but I believe 1-1/16 will be sufficiently accurate for poetry."

Today computers are standard tools for amateur and professional literary investigators alike. Shakespeare is both the most celebrated object of this effort and the most common. At Claremont McKenna College, in California, for example, two highly regarded faculty members have devoted years of their lives to a computer-based attempt to find out whether Shakespeare, rather than Francis Bacon or the Earl of Oxford or any of a myriad of others, wrote the plays and poems we associate with his name.

View This Story as a PDF

See this story as it appeared in the pages of The Atlantic magazine.

Open

As Babbage's venture into criticism foreshadowed, the marriage of computers and literature has been an uneasy one. At the mention of computers or statistics, many Shakespeareans and others in the literary establishment wrinkle their noses in distaste. To approach the glories of literature in this plodding way is misguided, they say, and misses the point in the same way as does the oft-cited remark that the human body is worth just a few dollars—the market value of the various chemicals of which it is composed. "This is just madness," says Ricardo Quinones, the chairman of the literature department at Claremont McKenna. "Why don't they simply read the plays?"

Rather than read, these literary sleuths prefer to count. Their strategy is straightforward. Most are in search of a statistical fingerprint, a reliable and objective mark of identity unique to a given author. Every writer will sooner or later reveal himself, they contend, by quirks of style that may be too subtle for the eye to note but are well within the computer's power to identify.

For a University of Chicago statistician named Ronald Thisted, the call to enter this quasi-literary enterprise came on a Sunday morning in December of 1985. Thisted had settled down with The New York Times Book Review and an article by Gary Taylor, a Shakespeare scholar, caught his eye. Taylor claimed that he had found a new poem by Shakespeare at Oxford's Bodleian Library. Among the many reasons Taylor advanced for believing in the authenticity of the poem, called "Shall I Die?," Thisted focused on one. "One of his arguments," Thisted says, "was that several words in the poem don't appear previously in Shakespeare. And that was evidence that Shakespeare wrote it. One's first reaction is, that's dumb. If Shakespeare didn't use these words, why would that be evidence that he wrote the poem?" But Taylor's article went on to explain that in practically everything he wrote, Shakespeare used words he hadn't used elsewhere. Thisted conceded the point in his own mind, but raised another objection. "If ALL the words in there were ones that Shakespeare had never used," he thought, "if it were in Sanskrit or something, you'd say, 'No way Shakespeare could have written this.' So there had to be about the right number of new words." That question—how many new words an authentic Shakespeare text should contain—was similar to one that Thisted himself had taken on a decade before. Together with the Stanford statistician Bradley Efron, then his graduate adviser, Thisted had published a paper that proposed a precise answer to the question "How many words did Shakespeare know but never use?" The question sounds ludicrous, like "How many New

Year's resolutions have I not yet made?" Nonetheless, Efron and Thisted managed to answer it. They found the crucial insight in a generation-old story, perhaps apocryphal, about an encounter between a mathematician and a butterfly collector.

R. A. Fisher, the statistical guru of his day, had been consulted by a butterfly hunter newly back from Malaysia. The naturalist had caught members of some species once or twice, other species several times, and some species time

and time again. Was it worth the expense, the butterfly collector asked, to go back to Malaysia for another season's trapping? Fisher recast the question as a mathematical problem. The collector knew how many species he had seen exactly once, exactly twice, and so on. Now, how many species were out there that he had yet to see? If the collector had many butterflies from each species he had seen, Fisher reasoned, then quite likely he had sampled all the species that were out there. Another hunting trip would be superfluous. But if he had only one or two representatives of most species, then there might be many species yet to find. It would be worth returning to Malaysia. Fisher devised a mathematical way to make that rough idea precise (and reportedly suggested another collecting trip). Efron and Thisted's question was essentially the same.

Where the naturalist had tramped through the rain forest in search of exotic butterflies, the mathematicians could scan Shakespeare in search of unusual words. By counting how many words he used exactly once, exactly twice, and so on, they would attempt to calculate how many words he knew but had yet to use.

RECOMMENDED READING

What Do Early KonMari Adopters' Homes Look Like Now?

JOE PINSKER

Neither Efron nor Thisted had imagined that their statistical sleight of hand could ever be put to a live test. No new work of Shakespeare's had been unearthed for decades. Now Taylor had given them their chance. A new Shakespeare poem, like a new butterfly-collecting trip to the jungle, should yield a certain number of new words, a certain number that Shakespeare had used once before, and so on. If Shakespeare did write "Shall I Die?," which has 429 words, according to the mathematicians' calculations it should have about seven words he never used elsewhere; it has nine. To Efron and Thisted's surprise, the number of words in the poem which Shakespeare had used once before also came close to matching their predictions, as did the number of twice-used words, all the way through to words he had used ninety-nine times before. The poem, which sounds nothing like Shakespeare, fit Shakespeare like a glove.

This is work that can suck up lives. One Defoe scholar, trying to pick out true

Defoe from a slew of anonymous and pseudonymous works, has pursued his quarry for twenty years, with no end in sight. A team trying to determine if the Book of Mormon was composed by ancient authors or by the nineteenth-century American Joseph Smith took 10,000 hours to produce a single essay. (The largely Mormon team of researchers concluded that Smith had not written the Book of Mormon. Confirmed samples of Smith's prose, the researchers argued, showed patterns of word usage different from those in the Book of Mormon.) Paper after paper begins with a trumpet fanfare and ends with a plaintive bleat. One writer, for instance, decided to determine whether Jonathan Swift or one of his contemporaries had written a particular article, by pigeonholing his words according to what part of speech they were. "The only positive conclusion from over a year of effort and the coding of over 40,000 words," she lamented, "is that a great deal of further study will be needed." (Swift himself had satirized, in Gulliver's Travels, a professor who had "employed all his Thoughts from his Youth" in making "the strictest Computation of the general Proportion there is in Books between the Numbers of Particles, Nouns, and Verbs, and other Parts of Speech.")

Despite the shortage of triumphs the field is growing, because more and more of the work can be assigned to electronic drudges. Scholars once had to count words by hand. Later they had the option of typing entire books into a computer, so that the machine could do the counting. Today computers are everywhere, and whole libraries of machine-readable texts are available. Software to do deluxe slicing and dicing is easy to obtain.

As a result, everything imaginable is being counted somewhere. Someone at this moment is tallying up commas or meticulously computing adjective-to-adverb ratios. But sophisticated tools don't automatically produce good work. A future Academy of Statistics and Style might take as its motto the warning that the Nobel laureate P. B. Medawar issued to his fellow scientists: "An experiment not worth doing is not worth doing well."

Among those least likely to be fazed by such pronouncements is a professor of political science at Claremont McKenna College named Ward Elliott. Elliott is an authority on voting rights, a cheerful eccentric, and, like his father before him, inclined to view the Earl of Oxford as the true author of Shakespeare's works. Four years ago Elliott recruited Robert Valenza, an expert programmer also on the Claremont McKenna faculty, and the two set to work on the authorship question.

This time the model would be not butterfly hunting but radar. Valenza had

spent considerable time devising mathematical procedures to find the patterns obscured by noisy and jumbled electronic signals. Adapted to Shakespeare, the idea was to go beyond counting various words, as many others had done, and see whether consistent patterns could be found in the way certain key words were used together. Two writers might use the words "blue" and "Green" equally often throughout a text, for example, but the writers could be distinguished if one always used them on the same page while the other never used them together.

This pattern-finding mathematics is widely used in physics and engineering, in deciphering television and radar signals, for example. Given a long list of words—not simply the "blue" and "Green" of the example, but dozens more—the computer can quickly tell how Shakespeare typically balanced those words. "You might have a pattern," Valenza says, "with a lot of 'love,' very little 'hate,' and a good deal of 'woe.'" A different writer might use the same words, and even use them at the same rates as Shakespeare, but the

configurations might be different. The result is that a given list of words produces a kind of voiceprint for each author.

Valenza and Elliott examined "common but not too common" words that Shakespeare used. To examine rare words, Valenza had reasoned, would be like trying to identify a voice from a whisper, and to examine common words would be to let the writer shout into your ear. The final, fifty-two-word list—with such miscellaneous entries as "about," "death," "desire," "secret," and "set"—was assembled by trial and error. It consisted of words with two key properties. In various works of Shakespeare's those words are used in patterns that yield the same voiceprint each time. And when other writers are tested, the same words yield voiceprints that are different from Shakespeare's.

The machinery in place, Valenza and Elliott began by testing Shakespeare's poetry against that of thirty other writers. Exciting results came quickly: The disputed "Shall I Die?" poem seemed not to be Shakespeare's after all. Three of the leading claimants to Shakespeare's work—Francis Bacon, Christopher Marlowe, and Sir Edward Dyer—were decisively ruled out. To Elliott's good-humored consternation, the test dealt just as harshly with the claims put forward on behalf of the Earl of Oxford. Worse was to follow. For even as this first round of tests ruled out the best-known Shakespeare candidates, it left a few surprising contenders. One possibility for the "real" Shakespeare: Queen

Elizabeth I. "That did it for our chance of appearing in Science," Elliott laments, "But it vastly increased our chance of getting into the National Enquirer." (To his dismay, Elliott did find himself in Science, not as the co-author of a weighty research paper but as the subject of a skeptical news brief with the headline "Did Queen Write Shakespeare's Sonnets?")

Valenza and Elliott have since conducted more-extensive tests that have ruled out Queen Elizabeth. But the mishap highlights a risk that is shared by all the number-crunching methods. "If the glass slipper doesn't fit, it's pretty good evidence that you're not Cinderella," Elliott points out. "But if it does fit, that doesn't prove that you are."

The risk of being fooled is least for someone who combines a deep knowledge of literature with some statistical insight. Donald Foster, a professor of English at Vassar College, fits that bill. Foster's scholarship is highly regarded. Soon after "Shall I Die?" was presented to the world, for example, he wrote a long

debunking essay that persuaded many readers that the poem was not Shakespeare's. In a more recent essay he consigned whole libraries of research to the scrap heap. Hundreds or thousands of articles have been written to explain the epigraph to Shake-speare's Sonnets, which begins, "To the onlie begetter of these insuing sonnets, Master W.H." Who was W.H.? Foster's solution to the mystery, which won him the Modern Language Association's Parker Prize, is that W.H. was...a typo. The publisher, who wrote the epigraph as a bit of flowery praise to honor Shakespeare, had intended to print "W.SH."

Those essays had nothing to do with statistics, but Foster has done some statistical sleuthing of his own, and he is well aware of the hazards. One scholar compared Shakespeare's plays with someone else's poems, for example, and concluded that Shakespeare used the present tense more than other writers do. Another compared Shakespeare with later writers and concluded that he used many four-letter words, whereas other writers used shorter words—forgetting that archaic words like "thou" and "hath" drive Shakespeare's average up. "There are strong and compelling reasons for avoiding this kind of research," Foster says, "because it's so difficult to anticipate all the pitfalls." But Foster himself has often given way to temptation. Like many Shakespeareans, he steers clear of the "authorship question," but he has looked into a pertinent mystery.

Shakespeare acted in his plays. But with two exceptions, we don't know what roles he took. Foster believes he has found a statistical way to gather that long-vanished knowledge. "It occurred to me," he says, "that Shakespeare may have been influenced in his writing by the parts he had memorized for performances and was reciting on a more or less daily basis." Last year Foster figured out a way to test that hunch. "The results," he says, "have been absolutely stunning."

"We started by using a concordance to type in all the words that Shakespeare used ten times or fewer," Foster says. These aren't exotic words, necessarily, just ones that don't crop up often in Shakespeare. Scholars have known for some time that these "rare" words tend to be clustered chronologically. Foster found that if two plays shared a considerable number of rare words, in the later play those words were scattered randomly among all the characters. In the earlier play, the shared words were not scattered. "In one role," Foster says, "there would be two to six times the expected number of rare words." There

stood Shakespeare: the words that Shakespeare the writer had at the tip of his pen were the ones he had been reciting as Shakespeare the actor.

If Foster is right, Shakespeare played Theseus in A Midsummer Nights Dream and "Chorus" in Henry V and Romeo and Juliet. In play after play the first character to come on stage and speak is the one that Foster's test identifies as Shakespeare: John Gower in Pericles, Bedford in Henry VI, Part I, Suffolk in Henry VI, Part II, and Warwick in Henry VI, Part III. And Foster's test picks out as Shakespeare's the two roles that we have seventeenth-century evidence he played: the ghost in Hamlet and Adam in As You Like It.

The theory can be tested in other ways. It never assigns to Shakespeare a role we know another actor took. The roles it does label as Shakespeare's all seem plausible—male characters rather than women or children. The test never runs in the wrong direction, with the unusual words scattered randomly in an early play and clustered in one role in a later play. On those occasions when

Foster's test indicates that Shakespeare played TWO roles in a given play—Gaunt and a gardener in Richard II, for example—the characters are never onstage together. Foster's theory passes another test. When Foster looks at the rare words that Hamlet shares with Macbeth, written a few years later, those words point to the ghost in Hamlet as Shakespeare's role. And if Foster looks at rare words that Hamlet shares with a different play also written a few years later—King Lear, for example—those shared words also pick out the ghost as Shakespeare's role.

Additional evidence has been uncovered. After Hamlet, the ghost's vocabulary exerted a strong influence on Shakespeare's writing and then tapered off. But Shakespeare's plays went in and out of production. When Hamlet was revived several years after its first staging, and Shakespeare was again playing the ghost, he began again to recycle the ghost's vocabulary.

It is a strange image, a computer fingering a ghost. But it is a sign of things to come. Eventually the prejudice against computers in literary studies will give way. "The walls are sure to crumble," Ward Elliott says, "just as they did in baseball and popular music....Some high-tech Jackie Robinson will score a lot of runs, and thereafter all the teams in the league will pursue the newest touch as ardently and piously as they now shrink from it."