MIND

How a Computer Program Helped Show J.K. Rowling wrote A Cuckoo's Calling

Author of the *Harry Potter* books has a distinct linguistic signature

By Patrick Juola on August 20, 2013

Credit: iStock/basti_90

“The man who wrote the note is a German. Do you note the peculiar construction of this sentence?” These were the words of Sherlock Holmes in “A Scandal in Bohemia,” analyzing a note from a client, unmasking the King of Bohemia incognito, and incidentally, establishing himself as a brilliant literary analyst. It is impossible to keep a secret from the legendary Sherlock Holmes, who can read an ocean from a drop of water. Just as the paper would have carried the marks of the royal fingers, to the skilled reader the writing carried the marks of the royal mind.

Fiction has recently become fact with the improving science of

stylometry, the study of writing style. In 1964, Frederick Mosteller and David Wallace published a three-year study of the distribution of common words in the *Federalist Papers* and showed that the writing style of Alexander Hamilton and James Madison differed in subtle ways. For example, only Madison used the word “whilst” (Hamilton used “while” instead). More subtly, while both Hamilton and Madison used the word “by,” Madison used it much more frequently, enough that you could guess who wrote which papers by looking at how frequently the word was used. Mosteller and Wallace took this work it its conclusion, and were able to show that certain “disputed” papers, claimed by both Hamilton and Madison, were overwhelmingly likely to have come from Madison’s pen. Today, computers can do this type of analysis in seconds, whether to uncover a case of murder-disguised-as-suicide, study an anonymous medieval poem, resolve disputes about authorial credit, or even provide political asylum for a refugee. In the last case, for example, a critic of a repressive foreign government claimed asylum on the basis of articles he had written and published on-line. The problem, though, was that the articles had been published anonymously. This wouldn’t necessarily stop a repressive secret service, in a place where mere suspicion is enough for imprisonment. But this technology was able to convince the immigration judge of his authorship of the documents in question, and hence to let him stay.

Over the past decade, I have developed a computer program to do this sort of analysis of writing style, based on literally millions of different features. This program will take a sample of writing and determine, on the basis of similarity, who among a set of authors was most likely to have written that sample. In July, I received an

email from a reporter for London's *Sunday Times* asking if I could help them solve a mystery. The reporter had received a tip that J. K. Rowling had secretly penned a novel under a pen name: *The Cuckoo's Calling*, by Robert Galbraith, who was described as a former member of the Royal Military police, and whose novel had grown "directly out of his own experiences and those of his military friends." The tip was at least plausible. Rowling and Galbraith had the same agent and editor. The book was unusually accomplished for a supposed first-time novelist. And Galbraith, a man who had ostensibly spent years in uniform, was surprisingly good at describing women's clothing. But hard evidence was still lacking. The reporter wanted to know what the computer program could determine.

Language use is a set of personal choices. For example, the English language provides a tremendous number of choices for words to describe something bigger-than-big, words such as "huge," "giant," "enormous," or "colossal." Writers can choose to express an idea with a few precise words or a bunch of common, general ones, and similarly to break a complicated idea — or not — into bite-sized simple sentences. We're not even conscious of many of these choices.

In a famous experiment, psychologists showed that people's memory for the general meaning of sentences was much better than their ability to recall a sentence word-for-word. For example, subjects who heard the sentence — "*The raccoons raced up the tree and the dogs raced around them"* — were asked a few minutes later if they had heard the sentence: *The raccoons raced up the tree and the dogs raced around it.* If you're reading quickly, you may not have noticed the minor change in the last word. Most subjects couldn't tell the difference, either. People don't pay much attention to these common little words as long as they understand the meaning of the sentence (the dogs are running around a tree with some raccoons in it), but the author's fingerprints are visible in the pronoun choice.

The program I developed, JGAAP (Java Graphical Authorship Attibution Program) does a mathematical analysis of the degree of similarity across a huge number of features, far too many for any human analyst to keep track of. Mosteller and Wallace, for example, looked at about thirty different words. JGAAP can keep track of every word in a set of encyclopedias. By looking at Galbraith's language choices, the program could quantify the degree of similarity between Rowling and Galbraith. If they were completely different, this could effectively rule out Rowling as an author and discredit the tip. If they were very alike, especially in comparison with other authors of the same type, it would show she was a likely author. While this wouldn't prove that Rowling had written it, it would be a strong form of objective evidence.

It is important to decide carefully what kinds of similarities to look at. Not all choices are created equal; some choices (such as word

length) are easier to notice, control, and change than others (such as the use of prepositions). It's often better to examine many different features than only a few, and to run many analyses to see if they agree. For this analysis, I chose four separate groups of features that have been shown to provide useful information about authorship. Just as importantly, they are also relatively independent of each other, so they provided cross-checks on each other. One variable that I used, for example, is the distribution of word lengths. Each novel has a lot of words, each word has a length, and so one can get a robust description that such-and-such percent of the words in this document have exactly so-many letters. I was able to get a measurement of similarity, with 0.0 being identity and progressively higher numbers being greater dissimilarity.

Another feature was the 100 most common words. What percentage of the document were "the," what were "of," and so on. This is again a rich data set that is easy to extract by computer. Finally, I ran two tests based on authorial vocabulary. The first was on the distribution of character 4-grams, groups of four adjacent characters. These could be words, parts of words (like four letters "nsid" that would be inside the word "inside") or even parts of two words (like the four letters "n th" as part of the phrase "in the"). I also ran on word bigrams, pairs of adjacent words (like "pairs of," "of adjacent," and "adjacent words") again a feature with a good track record. One advantage of this approach is unfortunately also a disadvantage. With thousands of features tracked, it's difficult to point to any small set of features and say "*these* are what make this like Rowling." Stylometry, like sports, is often a game of inches.

For this study, the reporter and I selected a Rowling novel and stories by three similar novelists (all British female crime novelists: Rowling's own *The Casual Vacancy*, Ruth Rendell's *The St. Zita Society*, P.D. James' *The Private Patient* and Val McDermid's *The Wire in the Blood*) to see which one was most similar to Galbraith. Across these four analyses, Rowling was the only writer to consistently match styles. Val McDermid, for example, used word pairs in a very similar way to Galbraith, but her use of long and short words was highly unlike Galbraith. Word length distribution was similar to Rowling or to James.

Interpreting these results can be tricky, but simple statistics can illustrate how tight this match is. First, all of the authors except for Rowling were clearly excluded by at least one test. Whoever the author of *Cuckoo* was, it wasn't Ruth Rendell. With four authors, a randomly chosen author would be equally likely to be closest to James as to McDermid, or just as likely to be distinct from Rendell as Rowling. If the author wasn't any of the four, she would be just as likely to be "close" to Galbraith (meaning one of the top two likely authors in the list) as "distant" (the third or fourth candidate). To put this another way, if Rowling had not written *Cuckoo*, she would have only a 50/50 shot of having similar word lengths. She would also have only a 50/50 chance of having similar word pairs, of having

similar character clusters, or similar common words. Only one writer in 16 would be "lucky" enough to have that similar a writing style to Galbraith's. If Rowling wasn't the author, then the tipster had only about a 6% chance of naming someone that consistently similar.

Did this "prove" Rowling's authorship? Of course not. Even DNA can't do that; a DNA match simply means that the person of interest or someone with similar genes, possibly a family member, was involved. Stylometry is much less reliable and accurate than DNA — after all, your DNA is constant and absolutely constant and unvarying throughout your life, but if two novels didn't vary at all, they'd be the same novel. All we really knew that this point was that it was either by Rowling herself, or by someone who wrote in a very similar style to Rowling. But this was enough information for the *Sunday Times* to approach her agent. On July 13, 2013, she admitted that *The Cuckoo's Calling* was her work, and that she had hoped, by publishing under a pen name, to get feedback without expectations.

This technology is clearly a double-edged sword. If Rowling can be identified by computational analysis, what about whistleblowers? Is anyone safe from the modern equivalent of Sherlock's all-seeing eye? For the moment, yes. The person who truly violated Rowling's privacy was not my computer or even the *Sunday Times* reporter, but the tipster who suggested the investigation in the first place. It's simply not feasible to look at every potential author to see who might have written a book; without old-fashioned detective work (and informants), the haystack is still large enough that needles can successfully hide.

Are you a scientist who specializes in neuroscience, cognitive science, or psychology? And have you read a recent peer-reviewed paper that you would like to write about? Please send suggestions to Mind Matters editor Gareth Cook, a Pulitzer prize-winning journalist and regular contributor to NewYorker.com. Gareth is also the series editor of Best American Infographics, and can be reached at garethideas AT gmail.com or Twitter @garethideas.

ABOUT THE AUTHOR(S)

Patrick Juola teaches computer science at Duquesne University, in Pittsburgh, PA. He received a Ph.D. in computer science from the University of Colorado and worked as a postdoc in Experimental Psychology at the University of Oxford. Aside from stylometry, his research interests include digital humanities and computer security.

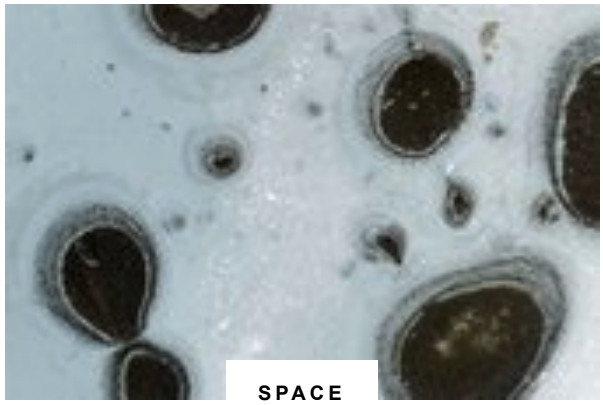
LATEST NEWS

Burning Man's Mathematical Underbelly

0 minute ago — Seth Stannard Cottrell and The Mathematical Intelligencer

South Africa Pushes Science to Improve Daily Life

1 hour ago — Sarah Wild and Nature magazine

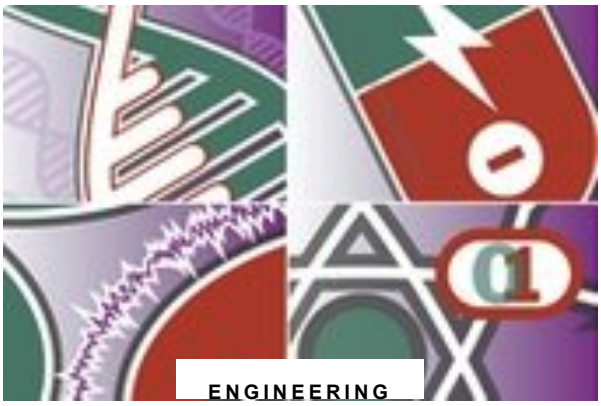


Searching for Life on Mars through the Lens of Greenland

2 hours ago — Marco Tedesco

The Environment's New Clothes: Biodegradable Textiles Grown from Live Organisms

2 hours ago — Erica Cirino



The Top 10 Emerging Technologies of 2018

2 hours ago — Mariette DiChristina and Bernard S. Meyerson

ENGINEERING

Top 10 Emerging Technologies of 2018

3 hours ago

NEWSLETTER

SIGN UP