phenomena.nationalgeographic.com

How Forensic Linguistics Outed J.K. Rowling (Not to Mention James Madison, Barack Obama, and the Rest of Us)

by Virginia Hughes • July 19, 2013 • 7 min read • original

Earlier this week, the UK's *Sunday Times* rocked the publishing world by revealing that Robert Galbraith, the first-time author of a new crime novel called *The Cuckoo's Calling*, is none other than J.K. Rowling, the superstar author of the *Harry Potter* series. Then the *New York Times* told the story of how the *Sunday Times's* arts editor, Richard Brooks, had figured it out.

One of Brooks's colleagues got an anonymous tip on Twitter claiming that Galbraith was Rowling. The tipster's Twitter account was then swiftly deleted. Before confronting the publisher with the question, Brooks's team did some web sleuthing. They found that the two authors shared the same publisher and agent. And, after consulting with two computer scientists, they discovered that *The Cuckoo's Calling* and Rowling's other books show striking linguistic similarities. Satisfied that the Twitter tipster was right, Brooks reached out to Rowling. Finally, on Saturday morning, as the *New York Times* reports, "he received a response from a Rowling spokeswoman, who said that she had 'decided to fess up'."

While the literary world was buzzing about whether that anonymous tipster was actually Rowling's publisher, Little, Brown and Company (it wasn't), I wanted to know how those computer scientists did their mysterious linguistic analyses. I called both of them yesterday and learned not only how the Rowling investigation worked, but about the fascinating world of forensic linguistics.

With computers and sophisticated statistical analyses, researchers are mining all sorts of famous texts for clues about their authors. Perhaps more surprising: They're are also mining not-so-famous texts, like blogs, tweets, Facebook updates

and even Amazon reviews for clues about people's lifestyles and buying habits. The whole idea is so amusingly ironic, isn't it? Writers choose words deliberately, to convey specific messages. But those same words, it turns out, carry personal information that we don't realize we're giving out.

"There's a kind of fascination with the thought that a computer sleuth can discover things that are hidden there in the text. Things about the style of the writing that the reader can't detect and the author can't do anything about, a kind of signature or DNA or fingerprint of the way they write," says Peter Millican of Oxford University, one of the experts consulted by the *Sunday Times*.

Cal Flyn, a reporter with the *Sunday Times*, sent email requests to Millican and to Patrick Juola, a computer scientist at Duquesne University in Pittsburgh. Flyn told them the hypothesis — that Galbraith was Rowling — and gave them the text of five books to test that hypothesis. Those books included *Cuckoo*, obviously, as well as a novel by Rowling called *The Casual Vacancy*. The other three were all, like *Cuckoo*, British crime novels: *The St. Zita Society* by Ruth Rendell, *The Private Patient* by P.D. James, and *The Wire in the Blood* by Val McDermid.

Juola ran each book (or, more precisely, the sequence of tens of thousands of words that make up a book) through a computer program that he and his students have been working on for more than 10 years, dubbed JGAAP. He compared *Cuckoo* to the other books using four different analyses, each focused on a different aspect of writing.

One of those tests, for example, compared all of the word pairings, or sets of adjacent words, in each book. "That's better than individual words in a lot of ways because it captures not just what you're talking about but also *how* you're talking about it," Juola says. This test could show, for example, the types of things an author describes as expensive: an expensive car, expensive clothes, expensive food, and so on. "It might be that this is a word that everyone uses, like expensive, but depending on what you're focusing on, it [conveys] a different idea."

Juola also ran a test that searched for "character n-grams", or sequences of adjacent characters. He focused on 4-grams, or four-letter sequences. For example, a search for the sequence "jump" would bring up not only jump, but jumps, jumped, and jumping. "That lets us look at concepts and related words without worrying about tense and conjugation," he says.

Those two tests turn up relatively rare words. But even a book's most common words — words like a, and, of, the — leave a hidden signature. So Juola's program also tallied the 100 most common words in each book and compared the small differences in frequency. One book might have used the word "the" six percent of the time, while another uses it only 4 percent.

Juola's final test completely separates a word from its meaning, by sorting words simply by their length. What fraction of a book is made of three-letter words, or eight-letter words? These distributions are fairly similar from book to book, but statistical analyses can dig into the subtle differences. And this particular test "was very characteristically Rowling," Juola says. "Word lengths was one of the strongest pieces of evidence that [*Cuckoo*] was Rowling."

It took Juola about an hour and a half to do all of these word-crunchings, and all four tests suggested that *Cuckoo* was more similar to Rowling's *Casual Vacancy* than the other books. And that's what he relayed back to Flyn. Still, though, he wasn't totally confident in the result. After all, he had no way of knowing whether the real author was somebody who wasn't in the comparison set of books who happened to write like Rowling does. "It could have been somebody who looked like her. That's the risk with any police line-up, too," he says.

Meanwhile, across the pond, Peter Millican was running a parallel Rowling investigation. After getting Flyn's email, Millican told her he needed more comparison data, so he ended up with an additional book from each of the four known authors (using *Harry Potter and the Deathly Hallows* as the second known Rowling book). He ran those eight books, plus *Cuckoo*, into his own linguistics software program, called Signature.

Signature includes a fancy statistical method called principal component analysis to compare all of the books on six features: word length, sentence length, paragraph length, letter frequency, punctuation frequency, and word usage.

Word frequency tests can be done in different ways. Juola, as I described, looked at word pairings and at the most common words. Another approach that can be quite definitive, Millican says, is a comparison of rare words. The classical example concerns the Federalist Papers, a series of essays written by Alexander Hamilton, James Madison, and John Jay during the creation of the U.S. Constitution. In 1963, researchers used word counts to determine the authorship of 12 of these essays that were written by either Madison or Hamilton. They found that Madison's

essays tended to use “whilst” and never “while”, and “on” rather than “upon”. Hamilton, in contrast, tended to use “while”, not “whilst”, and used “on” and “upon” at the same frequency. The 12 anonymous papers never used “while” and rarely used “upon”, pointing strongly to Madison as the author.

Millican found a few potentially distinctive words in his Rowling investigation. The other authors tended to use the words “course” (as in, of course), “someone” and “realized” a bit more than Rowling did. But the difference wasn’t statistically significant enough for Millican to run with it. So, like Juola, he turned to the most common words. Millican pulled out the 500 most common words in each book, and then went through and manually removed the words that were subject-specific, such as “Harry”, “wand”, and “police”.

Of all of the tests he can run with his program, Millican finds these word usage comparisons most compelling. “You end up with a graph, and on the graph it’s absolutely clear that *Cuckoo’s Calling* is lining up with *Harry Potter*. And it’s also clear that the Ruth Rendell books are close together, the Val McDermid books are close together, and so on,” he says. “It is identifying something objective that’s there. You can’t easily describe in English what it’s detecting, but it’s clearly detecting a similarity.”

On all of Millican’s tests, *Cuckoo* turned out to be most similar to a known Rowling book, and on four of them, both Rowling books were closer than any of the others. Millican got the files around 8pm on Friday night. Five hours later, he emailed the *Sunday Times*. “I said, ‘I’m pretty certain that if it’s one of these four authors, it’s Rowling.’”

This isn’t the first time that Millican has found himself in the middle of a high-profile authorship dispute. In the fall of 2008, just a couple of weeks before the U.S. presidential election, he got an email from the brother-in-law of a Republican congressman from Utah. He told him that they had used his Signature software (which is downloadable from his website) to show that Barack Obama’s book, *Dreams from my Father*, could have been written by Bill Ayers, a domestic terrorist. “They were planning to have a press conference in Washington to expose Obama one week before the election and got in touch with me,” Millican recalls, chuckling. “It was quite a strange situation to be in.”

Millican re-ran the analysis and definitively showed that *Dreams* was not, in fact, written by Ayers (you can read more about what he did here).

Juola told me some crazy stories, too. He once worked on a legal case in which a man had written a set of anonymous newspaper articles that were critical of a foreign government. He was facing deportation proceedings in the United States, and knew that if he was deported then the secret police in said foreign government would be waiting for him at the airport. Juola's analyses confirmed that the anonymous articles were, in fact, written by the man. And because of that, he was permitted to stay in the U.S. "We were able to establish his identity to the satisfaction of the judge," Juola says.

That story, he adds, shows how powerful this kind of science can be. "There are a lot of real controversies with real consequences for the people involved that are a lot more important than just, did this obscure novel get written by this particular famous author?"

The words of many of us, in fact, are probably being mined at this very moment. Some researchers, Juola told me, are working on analyzing product reviews left on websites like Amazon.com. These investigations could root out phony glowing reviews left by company representatives, for example, or reveal valuable demographic patterns.

"They might say, hmmm, that's funny, it looks like all of the women from the American West are rating our product a star and a half lower than men from the northeast, so obviously we need ot do some adjustment of our advertisements," he says. "Not many companies are going to admit to doing this kind of thing, but anytime you've got some sort of investigation going on, whether police or security clarance or a job application, one of the things you're going to look at is somebody's public profile on the web. Anything is fair game."

In fact, it was a good thing the original tipster of the Rowling news deleted his or her Twitter account, Juola says. "If we still had the account, we could have looked at the phrasings to see if it corresponded to anyone who works at the publishing house."

Original URL:
http://phenomena.nationalgeographic.com/2013/07/19/how-forensic-linguistics-outed-j-k-rowling-not-to-mention-james-madison-barack-obama-and-the-rest-of-us/