BOOKISH MATH

Statistical tests are unraveling knotty literary mysteries

BY ERICA KLARREICH

"The very thing!" exclaimed Professor Wogglebug, bounding into the air and upsetting his gold inkwell. "The very next idea!"

Devotees of Frank L. Baum's classic children's books would quickly recognize the above excerpt as the opening of the 15th book in the Oz series, *The Royal Book of Oz*. They might be harder pressed to say whether these lines were actually written by Baum. The book appeared with Baum's name on the cover in 1921, which was 2 years after Baum's death, and it was billed as the final work of the Royal Historian of Oz. For decades, however, fans and scholars have speculated that Ruth Plumly Thompson, who took over the series after Baum died, was the true author.

THE ROYAL BOOK OF OZ
By Ruth Plumly Thompson
Founded on and Continuing The Famous Oz Stories By L.Frank Baum
Illustrated by Jno. R. Neill

AUTHOR SWITCH — The modern cover of *The Royal Book of Oz* lists Ruth Plumly Thompson, not Frank L. Baum, as the author. A new mathematical analysis supports that attribution.

DOVER PUBLICATIONS

A few decades ago, literary detectives might have pinned their hopes of solving this mystery on finding the proverbial dusty manuscript in the attic trunk. Today, some scholars are tackling such problems with untraditional but more widely available tools: math formulas and computer programs.

Earlier this year, statistician José Binongo of the Collegiate School and Virginia Commonwealth University in Richmond published the results of statistical tests making a compelling case that Thompson wrote *The Royal Book of Oz*. Binongo's paper appeared in the spring *Chance*, in a special issue on stylometry—the science of measuring literary style.

Stylometry is now entering a golden era. In the past 15 years, researchers have developed an arsenal of mathematical tools, from statistical tests to artificial intelligence techniques, for use in determining authorship. They have started applying these tools to texts from a wide range of literary genres and time periods, including the *Federalist Papers*, Civil War letters, and Shakespeare's plays.

"We can now pretty accurately identify authorship—under the right conditions," says John Burrows, an emeritus English professor of the University of Newcastle in Australia.

What's more, the tremendous growth of computer power and electronic archives of literary texts is allowing stylometrists to carry out mathematical analyses on a scale previously unimaginable.

"Stylometry has a tremendous untapped potential," says Bernard Frischer, a classicist at the University of California, Los Angeles. He has used mathematical methods to study ancient Greek and Latin texts. "There are hundreds of insights waiting to be discovered by scholars who will take the time to learn statistics and computer programming," he says.

LITERARY FINGERPRINTS At first glance, it might appear that the way to pinpoint a writer's style is to study the rarest, most striking features of his or her writing. After all, it's the unexpected words and the unusual rhetorical flourishes that seem to mark a work as uniquely Shakespearean or Dickensian.

Yet the most venerable, commonly used approach of stylometrists does the opposite: It examines how writers use bread-and-butter words such as "to" and "with." Although this approach seems counterintuitive, it's based on sound logic.

"People's unconscious use of everyday words comes out with a certain stamp," says David Holmes, a stylometrist at the College of New Jersey in Ewing. Precisely because writers use these function words without thinking about them, they may offer more-reliable fingerprints of a writer's style than unusual words do.

"Rare words are noticeable words, which someone else might pick up or echo unconsciously," Burrows says. "It's much harder for someone to imitate my frequency pattern of 'but' and 'in.'"

In the early 1960s, statisticians Frederick Mosteller and David Wallace launched the use of function words to determine authorship. They analyzed the *Federalist Papers*, 85 essays published anonymously in 1787 and 1788 to persuade New Yorkers to adopt the new Constitution of the United States. Scholars have long known that Alexander Hamilton, James Madison, and John Jay wrote the essays, but both Hamilton and Madison claimed authorship of 12 of the papers.

To determine who wrote the disputed papers, Mosteller and Wallace compared word usage in other writings by Hamilton and by Madison. They found, for instance, that Hamilton used the word "upon" about 10 times as often as Madison did. Armed with 30 such distinguishing words, Mosteller and Wallace considered each disputed paper.

Mosteller and Wallace started out by assuming that for each paper, the probability was equal that Madison or Hamilton was the author. They then used the frequencies of the 30 words, one word at a time, to improve this probability estimate. They ultimately assigned all 12 disputed papers to Madison, a conclusion that dovetails with the historians' prevailing view.

Mosteller and Wallace's landmark study was the first convinc-