WHO WROTE THIS?: MODERN FORENSIC AUTHORSHIP ANALYSIS AS A MODEL FOR VALID FORENSIC SCIENCE

JANET AINSWORTH*
PATRICK JUOLA*

INTRODUCTION

On September 5, 2018, the *New York Times* published what might be, according to one commentator, "the most significant and consequential op-ed ever published."[1] The article, purportedly authored by a "senior official" within the Trump administration, contained a number of explosive assertions concerning the fitness of the President and claimed that the author, together with other officials in the administration, was "working diligently from within to frustrate parts of [Trump's] agenda and his worst inclinations."[2] Specifically, the article described Trump as "impetuous, adversarial, petty and ineffective,"[3] charged that "his impulsiveness results in half-baked, ill-informed, and occasionally reckless decisions,"[4] and stated that he "shows a preference for autocrats and dictators . . . [with] little appreciation for the ties that bind us to allied, like-minded nations."[5] As a result of the "instability"[6] displayed by the president, the author justified taking actions to thwart the president's agenda, arguing that since "the president continues to act in a manner that is detrimental to the health of our republic . . . our first duty is to this country . . . to preserve our democratic institutions . . . until he is out of office."[7]

In just a handful of days, the op-ed garnered more than 15,000 comments.[8] The commenters spanned the political spectrum, and raised numerous serious questions about the article and the actions of the author

* Janet Ainsworth is the John D. Eshelman Professor of Law at Seattle University. Patrick Juola is Professor of Computer Science and Director of the Evaluating Variations in Language Laboratory at Duquesne University. We would like to thank Brittney Adams for her excellent research assistance.

1. Michael J. Socolow, *The Times Would Have Been Crazy Not to Publish That Op-Ed*, POLITICO (Sept. 10, 2018), https://www.politico.com/magazine/story/2018/09/10/new-york-times-oped-defense-219745.
2. *I Am Part of the Resistance Inside the Trump Administration*, N.Y. TIMES (Sept. 5, 2018), https://www.nytimes.com/2018/09/05/trump-white-house-anonymous-resistance.html.
3. *Id.*
4. *Id.*
5. *Id.*
6. *Id.*
7. *Id.*
8. *Id.*

and his or her cohorts. These ranged from demands that the author and like-minded officials resign if they felt so strongly that they could not serve the aims of the president, to charges that they had engaged in a "deep state" coup d'état against an elected president, to pleas for them to provide Congress and the American people with evidence of President Trump's unfitness for office.[9] Indeed, President Trump himself took to Twitter to denounce the *New York Times* and the author, suggesting that one or both could be accused of treason, and urging Attorney General Jeff Sessions to order the Justice Department to investigate as a matter of national security.[10]

One question in particular seemed to command the public's attention: which of the officials in the Trump administration had written the article? Almost immediately, the speculation began. While some amateurs attempted to see if tell-tale indications of the author's identity could be gleaned from its wording, a major news organization approached a noted forensic authorship analyst to inquire whether he would be willing to analyze the text of the op-ed to determine its likely authorship.[11]

In a case such as this, even speculations can be dangerous. The consequences of a proposed attribution of authorship—from a professional source, in particular—could be both serious and extensive. Certainly, the identified author would likely be fired from his or her position in President Trump's administration. Sympathizers, suspected or proven, would also probably be dismissed. Depending on the extent of such a purge, the administration's functionality might be severely impacted. Further, public reaction to the identification and its aftermath could affect subsequent elections, or even completely derail the political careers of implicated individuals. With the stakes being so high, any authorship attribution would need to be correct, and demonstrably so. Only a scientifically validated analysis could suffice.

9. For a representative sample of these responses, see the more than 15,000 comments appended to the original op-ed article in the New York Times. *Id.*

10. Mark Landler & Katie Benner, *Trump Wants Attorney General to Investigate Source of Anonymous Times Op-Ed*, N.Y. TIMES (Sept. 7, 2018), https://www.nytimes.com/2018/09/07/us/politics/trump-investigation-times-op-ed.html.

11. *See, e.g.*, Nick Visser, *Omarosa Thinks She Knows Who Wrote that Anonymous New York Times Op-Ed*, HUFFINGTON POST (Sept. 12, 2018), https://www.huffingtonpost.com/entry/omarosa-identity-new-york-times-op-ed_us_5b98b241e4b0162f473214e5; Andrew Prokop, *Who Is The Senior Trump Official Who Wrote The New York Times Op-Ed?*, VOX (Sept. 6, 2018), https://www.vox.com/2018/9/6/17826830/new-york-times-trump-official-anonymous-oped. It was Patrick Juola who was contacted by a media source who we prefer to keep confidential. A copy of the email from the media source is on file with the authors and the *Washington University Law Review*.

I. WHY KNOWING WHO WROTE A QUESTIONED TEXT MATTERS

This episode demonstrates that the question "Who wrote this questioned text?" can have profound repercussions. Authorship attribution analysis—the forensic practice of examining texts to develop evidence as to the identity of the producer of a text of questioned provenance—is significant in many fields of inquiry. Literary scholars want answers to questions such as: Did Shakespeare write all the plays and sonnets attributed to him?[12] Or, in a more modern context, did the *Harry Potter* author J.K. Rowling also write the crime novel, *The Cuckoo's Calling*, under the pen name Robert Galbraith?[13] Journalists want answers to questions like: Who is the inventor of Bitcoin?[14] Historians are interested in questions such as: Who wrote the Jack the Ripper letters in Victorian England?[15] And: Which portions of the Federalist Papers were written by James Madison, which by Alexander Hamilton, and which by someone else?[16]

That last question might also be of interest to lawyers and legal scholars. In fact, for many issues in both criminal and civil contexts, authorship attribution can play a key role in arriving at the correct legal conclusion to the case. Here are just a few examples:

- Abby receives several emails threatening her life. The police want to know who wrote them.
- Bernie's new will changes his beneficiaries significantly. After Bernie's death, his former beneficiaries contest the will. Did

12. For modern authorship analyses of the corpus of Shakespeare's works, see Ward E.Y. Eliot & Robert J. Valenza, *And Then There Were None; Winnowing the Shakespeare Claimants*, 30 COMPUTERS & HUMAN. 191 (1996); *cf.* Joseph Rudman, *Non-traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats*, 5 J. EARLY MOD. STUD. 307 (2016) (agreeing generally with the methodology of Eliot and Valenza's work but raising some additional considerations and issues).

13. *See* Patrick Juola, *The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions*, 30 DIGITAL SCHOLARSHIP HUMAN. (Supplement 1) i100 (2015).

14. Matthew Herper, *Linguistic Analysis Says Newsweek Named the Wrong Man as Bitcoin's Creator*, FORBES (March 10, 2014), https://www.forbes.com/sites/matthewherper/2014/03/10/data-analysis-says-newsweek-named-the-wrong-man-as-bitcoins-creator/.

15. For a recent authorship analysis of the Jack the Ripper letters, see Andrea Nini, *An Authorship Analysis of the Jack the Ripper Letters*, 33 DIGITAL SCHOLARSHIP 621 (2018).

16. For an early example of analytic-based authorship attribution using Bayesian analysis on this question, see Frederick Mosteller & David L. Wallace, *Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers*, 58 J. AM. STAT. ASS'N 275 (1963) (attributing the questioned portions mainly to Madison); *cf.* David I. Holmes & Richard S. Forsyth, *The* Federalist *Revisited: New Directions in Authorship Attribution*, 10 LITERARY & LINGUISTIC COMPUTING 111 (1995) (applying several additional analytic techniques and largely corroborating Mosteller and Wallace).

Bernie actually compose it, or did one of the new beneficiaries do so?

- Charlene is found dead of a drug overdose, with an apparent suicide note found on her computer. Did Charlene write the suicide note, or was it written by Dana, who had a motive to kill her?[17]
- Defamatory online posts allege financial improprieties by Dean England. The dean suspects that the posts were written by ex-Professor Francis, who was denied tenure. Was the professor the author of the posts?
- Trade secrets of BizCo are revealed online. Those secrets are known to several people, including Gregory, who had signed a confidentiality agreement regarding trade secrets of BizCo. If Gregory exposed the trade secrets, BizCo intends to sue him for breach of that agreement. Who among the group of trade secret possessors actually revealed the secrets?[18]
- Henry and Isabel both submitted the same paper to a professor. Each one claims to have written the paper. Who did?[19]
- Jack signed a confession to a serious crime. He claims that the police added the incriminating parts of the "confession" to a non-incriminating part that he did write. Did Jack write the incriminating part of the confession?[20]
- Kerry is kidnapped and a ransom note sent to Kerry's wealthy parents. The ransom is paid, and Kerry released unharmed. The police suspect that the kidnapping was staged and that Kerry or

17. This hypothetical is based on a similar case analyzed by forensic stylometrician Carole Chaski, who showed that the roommate very likely authored the "suicide" note. Carole E. Chaski, *Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations*, 4 INT'L J. DIGITAL EVID. 1 (2005).

18. This hypothetical is similar to a case where the authorship of a set of emails was at issue, analyzed by Malcolm Coulthard. *See* Malcolm Coulthard, *On Admissible Linguistic Evidence*, 21 J. L. & POL'Y 441, 444–46 (2012).

19. Janet Ainsworth was consulted confidentially by a university professor about a situation in which two students each claimed to have written the same paper. Ultimately, one of the students confessed to have plagiarized the paper and withdrew from school, so no analysis in support of university disciplinary proceedings was required.

20. The example is based on the British prosecution of William Power for the alleged bombing of a pub. The forensic analysis of his purported confession, performed by Malcolm Coulthard, showed that Power likely was not the author of the incriminating portion of the "confession." Malcolm Coulthard, *Powerful Evidence for the Defence: An Exercise in Forensic Discourse Analysis*, *in* LANGUAGE AND THE LAW 414 (John Gibbons ed., 1994).

Kerry's partner wrote the ransom note. Did either of them write it?[21]

- Larry was fired after having sent racially offensive emails to his supervisor. He argued that several other employees had access to his computer and could have written the offensive emails. Did Larry actually write them?[22]
- A series of juvenile novels was authored first by Mack, and later by Nora, after Mack's death. However, it is disputed which of the two authored a novel written while Mack was ill. Which of the authors is due the royalties to the questioned novel?[23]
- Oliver, an asylum seeker, asserts that his life would be in danger if he were returned to his native country because of published articles critical of the government that he wrote under a pseudonym. If Oliver did write the articles, and that fact became known to the regime, he should be entitled to asylum in this country due to a well-founded fear of political persecution. So, did Oliver write the articles?[24]
- Police found a manifesto threatening a terrorist attack. The police want to identify the author to prevent the planned attack. Who wrote it?[25]
- Paul was a college chum of Quentin. He claims that he and Quentin jointly developed the idea behind a company that Quentin later started, and that Quentin sent him emails promising him a 50% share in the company in recognition of his contribution. Quentin denies both writing the emails and making the promise. Did Quentin write the promissory emails?[26]

---

21. The question of whether a purported ransom note is authentic or faked has been raised in the investigation of the murder of Jon Benét Ramsay, a case yet to be solved. *See, e.g.*, STEVEN THOMAS & DONALD A. DAVIS, JON BENÉT: INSIDE THE RAMSAY MURDER INVESTIGATION (2011).

22. This example comes from another real-world case. *See* Chaski, *supra* note 17.

23. A dispute paralleling these facts arose concerning the authorship of a book in the Wizard of Oz series. Jose Nilo G. Binongo, *Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution*, 16 CHANCE 9 (2003).

24. This is based on an actual legal dispute, analyzed by Juola, in which the asylum petition was eventually granted. Patrick Juola, *Stylometry and Immigration: A Case Study*, 21 J. L. & POL'Y 287 (2013).

25. For a discussion of this kind of problem, see Katerina T. Frantzi, *Computing Stylistics and Corpus Linguistics for Author Identification*, 1 INT'L J. INTERDISC. SOC. SCI. 69, 74 (2006).

26. This hypothetical is based on litigation against Mark Zuckerberg, where emails ostensibly by Zuckerberg admitted that Paul Ceglia was owed a large share of Facebook. A forensic stylistic expert, Gerald McMenamin, testified that the emails were not authored by Zuckerberg, although his conclusions were controversial among forensic linguists. For a recounting of the case and the controversy within the

These pseudonymous and sometimes- hypothetical cases illustrate just some of the many legal contexts in which authorship attribution questions arise. Forensic examination of the characteristics of the questioned texts can help shed light on such issues.

The forensic analysis of a text in order to determine its authorship has, for decades, been accomplished by forensic linguists. These experts are trained in the science of linguistics, and have often applied their specialized expertise in sociolinguistics, discourse analysis, and pragmatics. Although the range of applications for forensic linguistics is extremely wide,[27] this Article will focus only on the authorship attribution application. Herein, we maintain that reliability is an essential component of pattern comparison forensic practices, that testing for validity and measuring of error rates are the *Daubert* factors[28] that best guarantee reliability, and that the practices of authorship attribution can be an effective model of how to carry out this testing.

II. AUTHORSHIP ATTRIBUTION ANALYSIS AS A PATTERN-BASED FORENSIC SCIENCE: FOUNDATIONS FROM THE SCIENCE OF LINGUISTICS

All pattern comparison forensic practices begin with a premise: namely, that a creator's characteristics are reflected in his or her creation, such that patterns displayed in the creation can provide evidence regarding the identity of its maker. Consider, for example, forensic odontology, which begins with the assumption that the properties of one's teeth will be reflected in any of his or her resulting bitemarks.[29] Similarly, forensic authorship attribution begins with the linguistics-based premise that

---

field, see Lawrence M. Solan, *Intuition Versus Algorithm: The Case of Forensic Authorship Attribution*, 21 J. L. & POL'Y 551, 570–71 (2013).

27. *See, e.g.*, THE ROUTLEDGE HANDBOOK OF FORENSIC LINGUISTICS (Malcolm Coulthard & Alison Johnson eds., 2010); THE OXFORD HANDBOOK OF LANGUAGE AND LAW (Peter M. Tiersma & Lawrence M. Solan eds. 2012).

28. In *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, the Supreme Court set out a multi-factor analysis to govern judicial determinations of the admissibility of expert evidence, including (1) whether the scientific theory behind the evidence can be (and has been) tested; (2) whether the theory behind the evidence has been subjected to peer review; (3) the known or potential rate of error for the methods used to generate the evidence; (4) the existence and maintenance of standards controlling the technique's operation, and (5) whether the technique has achieved general acceptance in the relevant expert community. 509 U.S. 579, 593–97 (1993).

29. By using this example, the authors are in no way endorsing forensic evidence regarding bitemarks, which has been roundly criticized as inaccurate and invalid. *See, e.g.*, Michael J. Saks et al., *Forensic Bitemark Identification: Weak Foundations, Exaggerated Claims*, 3 J. L. & BIOSCIENCES 538 (2016). Forensic bitemark comparison practices have been so discredited that the PCAST report recommends that no further experiments be conducted in an attempt to validate this form of forensic practice since it inherently cannot be made more reliable. PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE & TECHNOLOGY, FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE COMPARISON METHODS 84–88 (2016) [hereinafter PCAST REPORT].

language users have individual preferences and habits that determine their use of language.

Just as a community's distinctive use of language can be said to constitute a dialect, an individual's distinctive use of language is said to be his or her idiolect.[30] According to Malcolm Coulthard—one of the most prominent forensic linguists[31]—a person's idiolect "will manifest itself in distinctive and cumulatively unique rule-governed choices for encoding meaning linguistically in written and spoken communications they produce."[32] Individual language users differ in both their linguistic repertoires (*i.e.*, the vocabulary and language patterns they know) and in the choices they make from those repertoires in producing speech and writings. In the event of competing possibilities, a person's choice of one word over another is not simply episodic and random; instead, says Coulthard, such a choice is based on an individual's personal "preference for selecting and combining these individual items in the production" of text or speech.[33] This concept of an individual idiolect makes possible the forensic practice of comparing a text where authorship is questioned with texts known to have been produced by a candidate for authorship—texts that, by the nature of their production, will display the candidate's idiolectical characteristics.

Of course, the corpus of texts known to be written by a candidate will never be a complete catalog of that person's idiolect. At most, they represent a sample from which only certain aspects of the idiolect can be deduced.[34] Note also that, although linguists have theoretically-grounded reasons to believe otherwise, it cannot be empirically proven that idiolects are individually unique.[35] Fortunately, for most authorship attribution problems the question is not whether, out of the seven and a half billion people in the world, the candidate is the only one who could have written the text. Instead, in most real-world applications, there is a closed and limited set of potential

30. *See generally* Malcolm Coulthard, *Author Identification, Idiolect, and Linguistic Uniqueness*, 25 APPLIED LINGUISTICS 431 (2004).

31. Coulthard is widely acknowledged to be a towering figure within the field of forensic linguistics. He founded the journal Forensic Linguistics, was the first president of the International Association of Forensic Linguists, and started the first graduate program in forensic linguistics in the English-speaking world at the University of Birmingham in the United Kingdom. He is the author of numerous books, book chapters, and articles in the field and has mentored countless scholars in forensic linguistics, including both authors here.

32. Coulthard, *supra* note 30.

33. *Id.*

34. *Id.*

35. That would entail analyzing the language patterns of every language user, an obvious impossibility. Empirically unfounded claims of individual uniqueness in many forensic sciences has been criticized as seriously misleading to fact-finders. *See, e.g.*, Simon A. Cole, *Forensics without Uniqueness, Conclusions without Individuation: The New Epistemology of Forensic Identification*, 8 L. PROBABILITY & RISK 233 (2009); Mark Page, Jane Taylor & Matt Blenkin, *Uniqueness in the Forensic Identification Sciences—Fact or Fiction?* 206 FORENSIC SCI. INT'L 2 (2011).

candidate authors (e.g., only the people who had access to the computer that the email came from), so that the analyst need only calculate the likelihood that one of the candidates created the text.

III. LINGUISTIC ANALYSIS IN AUTHORSHIP QUESTIONS IN LAW: A FORENSIC STYLISTIC APPROACH

Courtrooms around the world have had to deal with documents of questionable provenance. Linguists have often undertaken analyses of legally significant texts in order to draw conclusions about their authorship. One early example of this kind of linguistic detective work involved a kidnapping case in Illinois.[36] Therein, the ransom note included the sentence, "Put it in the green trash kan [sic] on the devil strip at the corner 18th and Carlson."[37] Not many English speakers would be familiar with the term "devil strip" as it is used in the ransom note. Presumably, the term was part of the author's idiolect; not only did he know the phrase, he also expected the note's recipient to understand it. So, who was the author (and presumptive kidnapper)? The Dictionary of American Regional English defines the term "devil strip" as "the strip of grass and trees between sidewalk and curb,"[38] and limits its use to the Akron area of Northeast Ohio. Someone aware of that fact could (in Sherlock Holmesian fashion) infer that the note writer shared the language practices of natives of that area. In this case, the role of Holmes was played by a sociolinguist and dialectologist named Roger Shuy, whose conclusion—that the ransom note writer was likely from Akron—led the police to investigate and ultimately convict the only one of the suspects hailing from that area.[39]

When looking for clues, linguistic analysts examine systematic language variation on many levels. These systematic language usage patterns were called "style markers," and analysis based on style markers became "forensic stylistics."[40] Patterns of punctuation, spacing, and spelling—particularly if they vary from standard English patterns—could be style markers reflecting an idiolect, as could word choices from among near-synonyms, and lexical choices derived from regional dialects or slang. At the sentence level, style markers may include grammatical choices, such as a preference for embedded clauses, a tendency to use parallel clause

36. Jack Hitt, *Words on Trial*, NEW YORKER (July 16, 2012), https://www.newyorker.com/magazine/2012/07/23/words-on-trial.

37. *Id.*

38. Roger W. Shuy, *DARE's Role in Linguistic Profiling*, 4 DARE NEWSLETTER 1 (2001).

39. *Id.*

40. *See generally* GERALD R. MCMENAMIN, FORENSIC STYLISTICS (1993); GERALD R. MCMENAMIN, FORENSIC LINGUISTICS: ADVANCES IN FORENSIC STYLISTICS (2002).

structures, the use or avoidance of contractions, or the use of what linguists would call "*that* complementizer deletion," *inter alia*.[41] Additionally, at the whole-text discourse level, style markers may include preferred narrative structures, levels of formality or informality that would be atypical for the type of text, and the use of irony, sarcasm, or hyperbole. At all levels, style markers can be assessed both qualitatively (*i.e.*, how unusual did they seem) and quantitatively (*i.e.*, how many times did they occur in the known text corpus, and how often in the questioned text).

For example, consider the case wherein forensic linguist Robert Leonard performed an analysis of two sets of letters featuring a number of peculiar style markers. One uncommon rhetorical device observed by Leonard in both letters—which he called "ironic repetition"—consisted of repetitions of a verb in consecutive sentences where the context of use was changed in each iteration so as to express irony by (1) altering the subject, and (2) shifting the complement of the verb.[42] Additionally, each of the two letters contained a pattern of contractions in which negatives were sometimes contracted—thus, "cannot" and "can't" were both used—but non-negatives were never contracted—for example, "I am" never was rendered as "I'm". These unusual patterns occurred in both the anonymous and identified letter, and helped Leonard to conclude that it was likely that a single author had composed both. As striking as these shared idiosyncrasies are *after* they've been pointed out, it is unlikely that someone untrained in linguistics would have spotted them or appreciated their significance. The linguistic evidence was key to resolving the police investigation; after the suspect was confronted with it, he confessed and was sentenced to prison.[43]

In their work, forensic linguists utilize their expertise in linguistics, dialectology, sociolinguistics, syntax, and discourse analysis. Their determinations about what features of a text should be considered style markers, and how significant those markers are in assessing authorship, are based on that science-grounded expertise. In the past, however, they had few or no methods with which to objectively test their analyses. Given the

41. "*That* complementizer deletion" refers to the choice to include or delete the word "that" in sentences like "I said I would go" instead of "I said that I would go." Both are grammatically available choices in standard American English, but an individual might tend to choose one option over the other on a reasonably consistent basis in that person's idiolect.

42. The stalker letter contained the lines, "I would have loved to have *found out*. A couple of days later she made sure my fiancé *found out*. She dumped me and then had an abortion." (emphasis added). In the serial killer letter—allegedly written by someone else—were the lines, "We had an affair . . . [and] she wanted to *break* it off. So I *broke* her neck!" (emphasis added). Leonard noted that this rhetorical device of ironic repetition was so unusual that experts in rhetoric did not readily have a label to apply to it. Robert Leonard, Juliane E.R. Ford & Tanya Karoli Christensen, *Forensic Linguistics: Applying the Science of Linguistics to Issues of the Law*, 45 HOFSTRA L. REV. 881, 887–88 (2017).

43. *Id.* at 887–89.

nearly limitless number of potential style markers,[44] developing authorship testing problems was thought to be impractical. The absence of validity testing did not, however, mean that analyses based on style makers were incorrect. Lack of proof of validity is not evidence of invalidity.[45]

The lawyer and linguist Lawrence Solan considered the question of how the legal system should judge the admissibility of forensic stylistic analyses conducted by linguists who had not subjected their methods to objective testing.[46] He contrasted two kinds of authorship attribution experts, calling one Lucy and the other Lacy.[47] Lucy is trained in using computer algorithmic analysis of textual features, has tested those algorithms on texts, and can supply a known error rate. Her methods provide correct answers in 88% of her tests, and she is working to improve the algorithms to increase the accuracy of her methods. Lacy, on the other hand, is trained in linguistics and uses style markers to analyze texts. Her experience in analyzing texts and her training in linguistics have led her to conclude that a set of 36 style markers can be used to determine whether a known author likely composed a questioned text. She has never tested the method on samples where the answer is already known, so she cannot provide an error rate—that said, she is confident that her analyses actually work. How then, Solan asks, should a judge determine the admissibility of testimony presented by Lucy versus Lacy? In considering that question, Solan relies in part on the fact that Lucy's methods have been subjected to validity tests. By examining the results of those tests, fact-finders can judge for themselves how to weigh her testimony. Thus, if the results of Lucy's tests exonerate the defendant, the fact-finder not only knows this conclusion by Lucy, but also knows that these tests, and the conclusions Lucy draws from them, have been shown to be fairly accurate (although not perfect).

By contrast, Lacy might actually have a higher accuracy rate than Lucy does. Lacy might only take on attribution cases in which the evidence overwhelmingly supports her conclusions. Conversely, Lacy's error rate may be enormous. Without validity testing, neither Lacy nor the factfinders weighing her testimony have any way of calculating the method's true error

44. *See* Joseph Rudman, *The State of Authorship Attribution Studies: Some Problems and Solutions*, 31 COMPUTERS & HUMAN. 351, 358 (1998) (positing a finite number of style markers, albeit more than a thousand).

45. David Faigman has made this point with respect to forensic practices more generally. He notes that a lack of empirical testing doesn't necessarily mean that a practice is without merit, only that it is not possible to determine a priori exactly what merit the practice might have. David L. Faigman, *Anecdotal Forensics, Phrenology, and Other Abject Lessons from the History of Science*, 59 HASTINGS L.J. 979, 985 (2008).

46. Solan, *supra* note 26.

47. *Id.* at 555–57, 564.

rate; thus, the jurors have no basis for judging whether her conclusions in these cases are likely to be right or wrong.

This problem was identified by a 2016 report on the use of forensic evidence in court, published by the President's Council of Advisors on Science and Technology (PCAST).[48] The report argued that "[f]oundational validity for a forensic-science method requires that it be shown, based on empirical studies, to be *repeatable, reproducible, and accurate*, at levels that have been measured and are appropriate to the intended application," and insisted that "neither experience, nor judgment, nor good professional practice (such as certification programs and accreditations programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability."[49]

Our contention is this: today, there are numerous methods of authorship attribution that can be shown to be foundationally valid. Indeed, empirical validity testing of these methods is currently being accomplished, with the results appearing in scholarly journals and magazines that are available to researchers, students, lawyers, and judges. Thus, the analysis of language for authorship has been shown to be "repeatable, reproducible, and accurate," in the language of the PCAST recommendations, and an active research community continues work on refinements to improve its reliability in both theory and practice. What follows is the story of those methods—forensic stylometry—and a summary of the research validating them.

IV. FORENSIC STYLOMETRY AS A TESTABLE AND TESTED METHOD OF AUTHORSHIP ATTRIBUTION ANALYSIS

A. Language can be Analyzed by its Objectively Identifiable Features

In some instances, the analysis of language can be highly subjective. Think about questions of literary interpretation, for example, where debates about the meaning and interpretation of a text are both acrimonious and commonplace.[50] High-level features of language (such as markers of

48. *See generally* PCAST REPORT, *supra* note 29. Forensic authorship attribution was not among the feature comparison practices examined by PCAST in this report.

49. *Id.* at 6. The PCAST report went on to caution that error rates cannot be deduced from individual casework but must be derived from analyses where the true answer was known. *Id*. at 53.

50. Literary theory and criticism incorporates a host of competing theoretical and ideological groundings that generate incommensurate readings of literature. For some recent attempts to summarize the various approaches to literary criticism, see, e.g., MARY KLAGES, LITERARY THEORY: A GUIDE FOR THE PERPLEXED (2006); MICHAEL RYAN, LITERARY THEORY: A PRACTICAL INTRODUCTION (2d. ed. 2007); TERRY EAGLETON, LITERARY THEORY: AN INTRODUCTION (3d. ed. 2008); PETER BARRY, BEGINNING THEORY: AN INTRODUCTION TO LITERARY AND CULTURAL THEORY (3d. ed. 2009); JONATHAN CULLER, LITERARY THEORY: A VERY SHORT INTRODUCTION (2d. ed. 2011); HANS

authorial intent or stance) are sufficiently subjective so that there can be substantial disagreement about how to assess the significance of a noticed feature, or even whether it is a feature at all. If one of the hallmarks of a reliable forensic science is that its analytic processes are objective and its results independently reproducible by other forensic experts, courts will be stymied in appropriately assessing techniques that are inherently dependent on the perspicuity of the analyst rather than by objective testing.

There is, however, a level of linguistic analysis that is straightforward and objective. Consider the following sentence: "THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG," and the following observations:

1. The example sentence is in the English language.
2. It contains nine words.
3. The word "THE" appears twice in it.
4. The word "THE" contains three letters.
5. The verb of this sentence is "JUMPS."
6. The subject of this sentence is "FOX."
7. It is in the present tense.
8. It is in the active voice.

These basic facts admit no debate. They can be established—repeatably, reproducibly, and accurately—by superficial examination by any competent reader, and, in fact, many of them can be determined algorithmically by a simple computer program. This type of observation can therefore provide a solid basis for further analysis if we want our analysis to be objective; that is, repeatable by the original tester and reproducible by any other tester looking at the text.

Consider now the following hypothesis: different people, having different vocabularies, will have different average word lengths in their written texts.[51] This suggests a simple, objective method for determining whether a disputed document, "*Q*," is more likely to have been written by Candidate Author 1 or Candidate Author 2. So long as we have documents *K1* and *K2*—which were written by the two candidates, respectively—we

BERTENS, LITERARY THEORY: THE BASICS (3d. ed. 2014); LOIS TYSON, CRITICAL THEORY TODAY: A USER-FRIENDLY GUIDE (3d. ed. 2014).

51. This observation is not unique to us, but dates back to the nineteenth century. *See, e.g.*, Letter from Augustus de Morgan to Rev. W. Heald (Aug. 18, 1851), *in* MEMOIRS OF AUGUSTUS DE MORGAN BY HIS WIFE SOPHIA ELIZABETH DE MORGAN WITH SELECTIONS FROM HIS LETTERS, 214, 215–16 (S. Elizabeth de Morgan ed., 2010) ("[T]ry to balance in your own mind the question whether the latter [text] does not deal in longer words than the former [text]. It has always run in my head that a little expenditure of money would settle questions of authorship this way. . . . Some of these days spurious writings will be detected by this test.").

can calculate the average length of the words used in each document. If the average word length of *Q* is closer to the average word length of *K1* than to that of *K2*, then *Q* and *K1* are more likely to share authorship.

This method is clearly objective, repeatable, and reproducible; whether or not it is *accurate* must be determined empirically. But this would not be unduly difficult to do. By testing a large number of document sets in which the identity of the author of *Q* is actually known, researchers can establish the probability that comparison of average word length correctly matches authors with texts.

As it happens, although more accurate methods are available, comparing the average word length in documents is somewhat accurate. As one of our authors, Patrick Juola, has discussed elsewhere:

> If you actually get a group of documents together and compare how different they are in average word length, you quickly learn two things. First, most people are average in word length, just as most people are average in height. Very few people actually write using loads of very long words, and few write with very small words, either. Second, you learn that average word length isn't necessarily stable for a given author. Writing a letter to your cousin will have a different vocabulary than a professional article to be published in *Nature*.[52]

Besides the average word length method, researchers have identified other objective and accurate means of determining authorship. Take this example: The fifteenth book in L. Frank Baum's *Oz* series (*The Royal Book of Oz*) was published after the Baum's death. At that time, scholars disputed whether or not the book was based on a substantial draft by Baum, and merely finished and copy-edited by Ruth Plumly Thompson, the woman hired to continue the series after Baum's death. Although the original cover credited Baum as author, some believed that the book was almost entirely Thompson's work. Linguist José Binongo conducted an authorship attribution analysis in this case, examining the individual frequencies of the fifty most common words instead of average word length.[53] Using a well-defined multivariate statistical procedure, he was able to show that Baum's use of these fifty words was clearly different from Thompson's, and that the writing style of the *Royal Book* was much more typical of Thompson's patterns of usage than Baum's. Specifically, Baum used words like *but*, *so*, *very*, *that*, and *which* much more frequently than Thompson, who used words like *up*, *on*, *back*, *out*, *over*, and *down* more often than Baum. In Binongo's words, the "stylistic gulf" separating these two authors "confirms

52. Patrick Juola, *Rowling and 'Galbraith': An Authorial Analysis*, LANGUAGE LOG (July 16, 2013, 7:35AM), http://languagelog.ldc.upenn.edu/nll/?p=5315.

53. Binongo, *supra* note 23.

. . . [that f]rom a statistical standpoint, this book is much more likely to have been written by Thompson than by Baum."[54]

Analyzing the usage of common words in a text to determine authorship has a long history,[55] and there is an emerging scholarly consensus that this kind of analysis is both accurate and precise.[56] Importantly, in conducting this method, linguists are more interested in identifying words with low semantic content (*i.e.*, pronouns, articles, and prepositions) than those with high semantic content (*i.e.*, nouns, verbs, and adjectives). This is because words of the latter type typically depend on, and are determined by, the topic of a text. For example, a set of articles by various authors about the Supreme Court will all tend to feature nouns like "court," "opinion," and "argument," verbs like "overrule" and "decide," and adjectives like "unconstitutional" and "binding." Obviously, the identification of those words—words of high semantic content—would not be useful in distinguishing one author from another. Further, if you were trying to determine the authorship of a text about, say, vacation activities, and your corpus of known texts by a candidate author included texts about constitutional law, you would not expect the high value semantic words from the constitutional law corpus to crop up in the vacation text, even if it were by the same author as the constitutional texts. With respect to such high semantic content words, genre and topic *matter*. On the other hand, researchers have consistently found that, regardless of what they are writing about, authors use the same set of low-semantic-content words.

Researchers have continued to explore which of numerous, objectively-measurable attributes of texts are most useful in conducting authorship attribution analyses. Some researchers have suggested that accuracy can be improved by using multiple separate analyses,[57] and have even proposed formal, step-by-step protocols to ensure that linguistic analyses can be replicated consistently by different forensic laboratories.[58]

---

54. *Id.*

55. *See, e.g.*, Mosteller, *supra* note 16 (analyzing the authorship of disputed parts of the Federalist papers).

56. For a concise summary of the research behind this consensus, see Efstathios Stamatatos, *A Survey of Modern Authorship Attribution Methods*, 60 J. AM. SOC'Y FOR INFO. SCI. & TECHNOLOGY 538, 540–41. (2009). *See also* Patrick Juola, *Authorship Attribution*, 1:3 FOUND. & TRENDS INFO. RETRIEVAL 233, 233–34 (2008); Moshe Koppel, Jonathan Schler & Shlomo Argamon, *Computational Methods in Authorship Attribution*, 60 J. ASS'N FOR INFO. SCI. & TECH. 9, 9–26 (2009).

57. Robert Bryll, Ricardo Gutierrez-Osuna & Francis Quek, *Attribute Bagging: Improving Accuracy of Classifier Ensembles by Using Random Feature Subsets*, 36:6 PATTERN RECOGNITION 1291, 1291–1302 (2003); Patrick Juola, *Authorship Attribution: What Mixture-of-experts Says We Don't Yet Know*, PROC. AM. ASS'N FOR CORPUS LINGUISTICS 233, 233–333 (2008).

58. Juola, *supra* note 13, at i100–i113.

B. A Case Study in Forensic Stylometry: How Authorship Analysis Revealed Who Penned The Cuckoo's Calling[59]

A highly-publicized example of how forensic stylometric methods can unmask an author occurred when it was rumored that *The Cuckoo's Calling* ("*Cuckoo*")—a detective novel published under the name Robert Galbraith—may have in fact been written by J.K. Rowling, author of the best-selling *Harry Potter* series. In July of 2013, a tweet claimed that 'Robert Galbraith' was actually a pen name for Rowling. As the rumor circulated, the Sunday Times of London contacted Patrick Juola and asked him to determine the likelihood that Rowling was actually the book's author.

In order to begin his analysis, Juola needed a corpus of texts written by Rowling to compare with *Cuckoo*. He used her novel, *The Casual Vacancy* ("*Vacancy*"), as the known text. Next, he compiled a set of texts from "distractor" authors to see how similarities between *Cuckoo* and *Vacancy* compared with similarities between *Cuckoo* and texts by the distractor authors. Juola selected distractor authors demographically similar to Rowling (*i.e.*, British female authors in Rowling's age demographic[60]), and used texts of those authors in the same genre as *Cuckoo*—that is, contemporary crime novels.[61] Finally, with the help of a computer program, Juola ran separate analyses to identify and compare four objective characteristics of *Cuckoo*, *Vacancy*, and the three "distractor" texts: (1) word length, (2) character 4-grams,[62] (3) word pairs, and (4) the 100 most frequently used words (which as an objective text feature has been shown to have discriminatory power).[63]

For each analysis, Juola plotted the similarities between *Cuckoo* and the four comparison texts. For example, the use of word pairs in *Cuckoo* was most similar to Rowling's uses, with McDermid a distant second. After each test, the author was judged not to be a plausible candidate author if his or

59. For a more detailed discussion of this case study and three other cases as well, see Juola, *supra* note 13.

60. Juola selected authors similar to Rowling in age, gender, and nationality to reduce the possibility that his analysis might be measuring one of those factors rather than authorship per se. He acknowledges that this level of similarity in "distractor" candidates may not have been necessary. *Id.* at i106. Holding those factors constant does, however, make his conclusions more compelling by eliminating that potential explanation for similarity or difference.

61. The "distractor" texts were Ruth Rendell's *The St. Zita Society*, P.D. James's *The Private Patient*, and Val McDermid's *The Wire in the Blood*.

62. A character 4-gram analysis is a specific type of n-gram analysis. N-gram analysis refers to chunking a text into consecutive words or characters in strings that are N-units long. For example, a character 4-gram analysis would break text into consecutive strings of four characters—letters or spaces. This analysis of the phrase "I LOVE YOU" would generate (1) "I space L O," (2) "space L O V," (3) "L O V E," (4) "O V E space," etc.

63. *See supra* note 60 and accompanying text.

her work was not among the top two matches against *Cuckoo*. The results were clear: "Of the four authors, Rowling, and only Rowling, was not eliminated by at least one analysis."[64] Moreover, because each of the analyses examined an independent variable in the text, it was possible to calculate the chances that the results were merely a random match. By the same token, if you knew the accuracy of the tests based on previous testing, you could calculate the chances of a false rejection of the true author. Indeed, if this had been a court case turning on the authorship of *Cuckoo*, Juola could have done more than merely testify about his conclusion; he also could have given the fact-finder an empirically-tested error rate for his methodology, providing a basis for judging the likelihood that he had falsely attributed the book, or falsely rejected the true author.[65]

V. ADVANCES IN AUTHORSHIP ATTRIBUTION ANALYSIS THROUGH VALIDITY TESTING: DEVELOPING THE NORMS OF A RESEARCH CULTURE

The key to validating the science behind authorship attribution has been the development of accuracy benchmarks through the use of shared evaluation corpora on which practitioners can test their methodologies.[66] These corpora consist of document sets with known "ground truths" about their authorship. The actual authorship of the documents in question is typically withheld from evaluation participants. In other words, participants testing the method in question are given documents whose authors are unknown to them at the time when they perform their analysis. Among the earliest of these tests was the 2004 Ad-hoc Authorship Attribution Competition.[67] This competition provided participants with thirteen sets of authorship attribution problems across a variety of languages and genres. For example, Problem A involved fixed-topic essays in Modern English. The languages used included not only Modern English, but also Middle English, French, Serbian-Slavonic, Latin, and Dutch. Genres included school essays, novels, plays, personal letters, meeting transcripts, historical writings, and poems. The results of this competition are a matter of public record, and have helped enable other scholars to evaluate the accuracy of any specific method. For example, the best performing method on Problem

64. Juola, *supra* note 13, at i108. Rowling did acknowledge being the author after Juola's analysis was made public. *Id.* at i111.

65. *Id.* at i108.

66. Corpora, the plural of corpus, refers to a collection of texts that can be data-mined for significant features.

67. *See* Juola, *supra* note 56, at 288.

A correctly identified the authors of eleven of the thirteen samples.[68] Someone guessing randomly would be expected to get only one correct, so this shows very clearly that this particular algorithm performs well on this specific kind of data.[69] This testing method gave us data illuminating the kinds of algorithms that were most accurate in solving a variety of authorship attribution problems.

Comparative evaluation of analyses continues in the field of computationally-assisted authorship attribution, most notably with the Plagiarism Action Network series of workshops.[70] The organizers of these workshops prepare the test corpora, and participants apply their analytic skills to determine the actual authorship of the documents. Once all participants have submitted their conclusions, the true answers are revealed, and the accuracy of each performer becomes a matter of public record.

For example, the PAN-2013[71] corpus contained texts in three languages, including English. The English portion consisted of fragments of computer science textbooks, and involved thirty "problems," each of which consisted of two-to-six text fragments written by a single author, and one fragment that may or may not have been by that same author.[72] The corpus was balanced so that fifteen of the problems were written by the same author and fifteen of the problems were written by different authors. Random guessing would thus provide a baseline of 50% accuracy. Eighteen teams submitted solutions to the English language problems, although not all participants

68. Vlado Keselj et al., *N-Gram-Based Author Profiles for Authorship Attribution*, 3 PROC. CONF. PAC. ASS'N FOR COMPUTATIONAL LINGUISTICS 255, 255–64 (2003).

69. Using the binomial theorem, statistical analysis confirms that this is a highly significant result ($p < 0.000001$). This also suggests that this particular method has an error rate of roughly 15% (2/13) under these exact conditions.

70. *See* Patrick Juola, *An Overview of the Traditional Authorship Attribution Subtask*, CLEF ONLINE WORKING NOTES/LABS/WORKSHOP (2012), http://ims-sites.dei.unipd.it/documents/71612/155385/CLEF2012wn-PAN-Juola2012.pdf; Patrick Juola & Efstathios Stamatatos, *Overview of the Author Identification Task at PAN 2013*, CLEF ONLINE WORKING NOTES/LABS/WORKSHOP (2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-JuolaEt2013.pdf; Efstathios Stamatatos, et al., *Overview of the Author Identification Task at PAN 2014*, CLEF ONLINE WORKING NOTES/LABS/WORKSHOP (2014), http://www.uni-weimar.de/medien/webis/events/pan-14/pan14-papers-final/pan14-authorship-verification/stamatatos14-overview.pdf#page=9; Efstathios Stamatatos, et al., *Overview of the PAN/CLEF 2015 Evaluation Lab*, INTERNATIONAL CONFERENCE OF THE CROSS-LANGUAGE EVALUATION FORUM FOR EUROPEAN LANGUAGES 518 (2015); Pablo Rosso et al., *Overview of PAN'16*, INTERNATIONAL CONFERENCE OF THE CROSS-LANGUAGE EVALUATION FORUM FOR EUROPEAN LANGUAGES (2016) https://www.uni-weimar.de/medien/webis/publications/papers/stein_2016i.pdf.

71. *See* Patrick Juola & Efstathios Stamatatos, *Overview of the Author Identification Task at PAN 2013*, CLEF ONLINE WORKING NOTES/LABS/WORKSHOP (2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-JuolaEt2013.pdf (3 et seq. describes the training documents; 9 et seq. describes the individual results).

72. Technically speaking, this task—where the question posed is "did this particular person write this particular document?" and where the answer is "yes" or "no" —is called "authorship verification." This is the authorship attribution task posed in the hypothetical case of Professor Francis, *supra* p. 4, or the Rowling case as described, *supra* p. 15. Authorship verification is a much harder problem than the more typical one of determining which of a pool of candidate authors wrote a particular document.

submitted solutions to all problems. As expected, individual results varied. While two teams failed to achieve the 50% baseline, eight teams achieved 70% accuracy or better, and two hit the 80% mark. Notably, combining all eighteen participants in a mixture-of-experts model yielded 86.7% accuracy. These results suggest two methods that should probably be avoided—the methods that undershot a chance score—eight methods that achieve significantly good accuracy, and a method of combining methods that is better still. In other words, the PAN-2013 workshop has provided accuracy norms, well-understood protocols to achieve those norms, and the measurements needed to assess the weight to be given to any particular analysis.

Even more importantly, the 2014 version of this conference was able to compare performance of the 2013 "winner" against the new submissions. The 2014 competition offered not one but two corpora: one of essays, another of novels. Of the thirteen participants in 2014, eleven outperformed the 2013 winner on essays, and all thirteen outperformed on the novels corpora.[73] This suggests not only that authorship attribution can be done with accuracy, but that public evaluations and objective benchmarks for accuracy can lead to substantially improved performance over a single year. Sharing information about which methods work best for which kinds of problems enables the entire field to build collectively on the research proffered by the community. Good methods are refined, while poorer methods are dropped by the wayside. This is the hallmark of what Thomas Kuhn calls a "mature science".[74]

How accurate are modern authorship attribution methods? As long ago as 2004, Juola showed that 90% accuracy was easily achievable on a sample of letters from the Ad-hoc Authorship Attribution Competition. More recently, Eder demonstrated that 80% accuracy could be achieved on a problem despite a much larger set of distractor authors—seventeen in all.[75] In 2013, the PAN/CLEF results showed nearly 90% accuracy on an authorship verification problem with no distractor authors at all. While these accuracy rates are less than ideal, they are still light years ahead of the accuracy rates achieved in most other pattern comparison forensic practices.[76]

---

73. The 2014 and later competitions used a more mathematically complex evaluation scheme. We have omitted the numerical details for simplicity.

74. *See* Juola, *supra* note 13, at i101.

75. Maciej Eder, *Does Size Matter? Authorship Attribution, Small Samples, Big Problem*, 30 DIGITAL SCHOLARSHIP HUMAN. 167, 167–82 (2013).

76. *See, e.g.*, PCAST REPORT, *supra* note 29, at 75–83 (summarizing the current state of forensic practices in testing complex admixtures of DNA), 83–87 (bitemarks), 104–14 (firearms ballistics), 114–17 (shoeprints), 117–21 (microscopic hair comparison). *See also* Bradford T. Ulery et al., *Understanding*

So far, almost all of the participants in PAN-like evaluations have been analysts using computational methods that utilize computer algorithms, but these testing sessions are not limited to researchers using computerized methods. Indeed, invitations to participate in these evaluations have been extended to forensic linguists that use more traditional methods. As more traditionally-trained forensic linguists come to recognize the importance of validation testing, we expect to see their increased involvement in future validation testing contests and workshops. In 2017, a particularly important authorship attribution workshop was held in conjunction with the biennial meeting of the International Association of Forensic Linguists. At this so-called "Forensic Linguistics Dojo", traditionally trained forensic linguists (as well as computational specialists using algorithms) were asked to analyze ten sets of documents and determine authorship.[77] The significance of this event cannot be overstated. The International Association of Forensic Linguists ("IAFL") serves as the disciplinary home for trained linguists operating in a variety of forensic contexts, including authorship analysis.[78] The group's choice to embrace validation testing and training is a clear indication that—despite the historical disciplinary divide between forensic stylistic analysts and computational analysts—practitioners in the field believe a touchstone for the reliability of forensic linguistic evidence is both necessary and desirable.

A key contribution of stylometric research to forensic practice, then, is its normative commitment to validation testing. By bringing experts together in a cooperative environment, analysts can calculate the accuracy of their methods in a testing environment that eliminates confirmation bias, establishes accuracy rates associated with specific techniques, and validates a continuously increasing standard of practice. Furthermore, the "open source" nature of these competitions makes it easy for researchers both to learn from each other, and spread knowledge among the authorship attribution community of methodological improvements. As a result, in

the Sufficiency of Information for Latent Fingerprint Value Determinations, 230 FORENSIC SCI. INT'L 99 (2013) (discussing problems of repeatability of analysis by individual fingerprint analysts and problems of reliability due to inconsistency between analyses done by multiple fingerprint analysts.)

77. As of this writing, the results have not yet been released.

78. Many of the leading figures in the IAFL have in recent years stressed the importance of validation testing in the authorship analysis context. In fact, three past presidents of the organization are on record as supporting validation testing and increased cooperation between qualitative and algorithm-driven forensic authorship analysts. In his 2012 Presidential address to the IAFL biennial meeting, Ron Butters spoke of the need for validation testing in forensic linguistics, calling it an ethical issue for testifying linguists. *See* Solan, *supra* note 26, at 554–55. Lawrence Solan organized a conference in 2012, calling for increased validation testing, particularly in forensic stylistics. *Id*. Malcolm Coulthard, a founder of the IAFL, spoke at the 2017 IAFL conference about his belief that future authorship analysis would need to combine computational and stylistic approaches to deliver the greatest accuracy in results. *Id.*

addition to saying that authorship analysis methods are more accurate and reliable today than they were a decade ago, we can specify what kinds of techniques are best for particular problems and quantify both the accuracy rates and the improvement in their accuracy rates.

VI. AUTHORSHIP ATTRIBUTION PRACTICES AS A MODEL FOR OTHER FORENSIC PRACTICES

A. The Crisis in Forensic Science: Unreliable Evidence is too often Admitted in Court

Science-based evidence is essential to a fair and accurate justice system. It promises to aid in accurately resolving issues in litigation that would otherwise be unprovable. That promise, however, is unfulfilled when proffered expert testimony turns out to be unreliable or, even worse, baseless. For decades, scholars have decried the admission into court of so-called "junk science".[79] Forensic identification evidence has been the target of particular criticism, with one scholar calling forensic pattern comparison sciences "contenders for the shoddiest science offered to the courts."[80]

In criminal cases, the admission of unreliable scientific evidence may result in tragic miscarriages of justice. The Exoneration Registry's list of criminal cases in which persons were erroneously convicted of crimes provides a hint of the extent of these costs, and includes information about the factors that contributed to each erroneous conviction.[81] Although the true number of cases in which flawed scientific evidence led to a miscarriage of justice will never be known, among the registry's documented erroneous convictions, about 20% involved the admission of inaccurate or unreliable forensic evidence.[82] Many of these cases involved innocent persons who spent decades imprisoned before being exonerated. A shocking number were awaiting execution. We will never know how many

79. *See, e.g.*, Randolph N. Jonakait, *The Meaning of Daubert and What That Means for Forensic Science*, 15 CARDOZO L. REV. 2103 (1993); Paul C. Giannelli, *Junk Science: The Criminal Cases*, 84 J. CRIM. L. & CRIMINOLOGY 105 (1993); Craig N. Cooley, *Reforming the Forensic Science Community to Avert the Ultimate Injustice*, 15 STAN. L. & POL'Y REV. 381 (2004); Erin Murphy, *The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence*, 95 CALIF. L. REV. 721 (2007).

80. Michael J. Saks, *Banishing* Ipse Dixit*: The Impact of* Kumho Tire *on Forensic Identification Science*, 57 WASH. & LEE L. REV. 879, 879 (2000).

81. *See* NAT'L REGISTRY OF EXONERATIONS, https://www.law.umich.edu/special/exoneration/Pages/about.aspx (last visited Feb. 26, 2019).

82. As of this writing, of the 2169 documented exonerations in the National Registry of Exonerations, 398 of those cases list unreliable forensic evidence as a sole or contributing factor in the wrongful conviction. *Id.*

innocent people convicted on the basis of unreliable forensic evidence were executed.[83]

Inaccurate sciences can have an impact in civil cases as well as in criminal ones.[84] When unreliable scientific testimony is admitted into the court and considered by judges or juries—and, equally, when reliable scientific evidence is excluded—the result is often a denial of justice. In numerous countries around the world, the problem of unreliable forensic science has been increasingly recognized as a major issue in both civil and criminal matters.[85]

Concern about the reliability of forensic evidence is not limited to legal scholars. In 2009, the National Academy of Sciences (NAS) concluded a four-year study on the state of forensic science in the United States.[86] The NAS commissioned a committee of scientists, judges, legal scholars, and forensic practitioners to conduct a searching examination of forensic evidence practices. What they found was an alarming lack of scientific validation for almost all forensic identification practices, notwithstanding practitioners' courtroom testimony claiming to match evidence to the person or source that produced it.[87] Indeed, experts were testifying in court "without any meaningful scientific validation, determination of error rates, or reliability testing."[88] Even more troubling, they were allowed to make probabilistic claims about matching patterns they found *despite* the lack of any statistical foundation for those claims.[89] The NAS's report exposed the

83. Unfortunately, we have no way of knowing how many innocent people have been executed because they lacked the opportunity to be exonerated. One executed person who was likely innocent is Cameron Todd Willingham. He was convicted of arson-murder on the testimony of a forensic arson expert that the fire in question was intentionally set. Today's understanding of arson science instead shows that the arson "signatures" in that case are typical of accidental fires. Willingham was executed in 2004, despite the fact that evidence had by then come to light debunking the testimony that convicted him. *But see* Paul L. Giannelli, *Junk Science and the Execution of an Innocent Man*, 7 N.Y.U. J. L. & LIBERTY 221 (2013) (describing the compelling evidence that now-debunked arson evidence led to the execution of Cameron Todd Willingham.)

84. Empirical examination of trial court rulings has shown that judges are generally more stringent regarding admissibility in civil cases when the evidence is offered by plaintiffs. Jennifer L. Groscup et al., *The Effects of* Daubert *on the Admissibility of Expert Testimony in State and Federal Criminal Cases*, 8 PSYCHOL. PUB. POL'Y & L. 339, 346 (2002).

85. *See, e.g.*, David E. Bernstein, *Junk Science in the United States and the Commonwealth*, 21 YALE J. INT'L L. 123 (1996); Carole McCartney & Clive Walker, *Forensic Identification and Miscarriages of Justice in England and Wales*, *in* ADVANCES IN FORENSIC HUMAN IDENTIFICATION 391 (Xanthe Mallett, Teri Blythe & Rachel Berry eds., 2010); Christophe Champod & Joselle Vuille, *Scientific Evidence in Europe—Admissibility, Evaluation, and Equality of Arms*, 9 INT'L COMMENT. ON EVIDENCE 1 (2011).

86. NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMY OF SCIENCES, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD (2009). [hereinafter NAS REPORT]. Forensic authorship attribution was not discussed within this report.

87. *Id.* at 47.

88. *Id.* at 108.

89. *Id.* at 188.

magnitude of the problem—almost none of the forensic evidence being routinely admitted into trials had been adequately validated—and called for a series of sweeping institutional reforms to improve the accuracy of forensic sciences and ensure that forensic evidence could be appropriately assessed when admitted in court.[90]

Seven years later, another comprehensive report—this one authored by PCAST—assessed the scientific foundations of a number of feature comparison forensic practices, including forensic comparison of latent fingerprints, bitemarks, toolmarks, shoe prints, hair taken from a crime scene, and firearms identification.[91] What they found was even more troubling than the conclusions of the NAS report. While the NAS focused on the institutional and resource challenges to creating reliable forensic science, PCAST looked in detail at the substance of what forensic experts asserted that they could do, and found many of the examined forensic fields to be seriously unreliable.[92] For example, after reviewing the literature on bitemark identification, the PCAST report recommended that no further research in that field be conducted, as it was extremely unlikely that bitemark forensics could ever provide reliable evidence on the source of purported bitemarks.[93] Their overall recommendation was that forensic feature-comparison fields should not be relied upon in court unless they could produce empirical testing showing the forensic method's error rate; that is, how often the method incorrectly concludes that samples from different sources came from the same source (a false positive) and how often the method incorrectly concludes that samples from the same source came from different sources (a false negative).[94]

B. Authorship Attribution Methods as a Model for Reliable Forensic Science Practices

Several characteristics of authorship attribution distinguish it from the types of forensic science criticized by NAS and PCAST. First, linguistic analysis is objective in a way that other feature comparison fields, such as bitemark and tire impression analysis, have not been.[95] Language is composed of a small set of elements that can be objectively identified. In a printed document, there no subjectivity in determining whether a given

90. *Id.* at 188–192, 213–216, 237–240.
91. PCAST REPORT, *supra* note 29.
92. The PCAST committee reviewed more than 2,000 published articles and took extensive written comments from interested parties in assessing feature-comparison forensic practices. *Id.* at 2.
93. *Id.* at 84–87.
94. *Id.* at 5–6, 65.
95. *Id.*

letter is an 'e' or an 'a.' Likewise, the way linguistic elements combine to make words and sentences is equally objective and countable. For this reason, it is possible to conduct linguistic analyses without relying on the intuitions of a particular expert, which can be neither validated nor replicated in future cases. Indeed, with forensic stylometry, such analyses can be consistently performed by any analyst anywhere, and can even be automated.[96]

A further strength of forensic authorship analysis is the existence of several well-curated collections of natural language that provide a baseline for quantitative assertions about linguistic features in texts. Consider, for example, a questioned text and a known text, both of which contain a specific word or phrase. If it can be shown that this expression is particularly rare, the fact that it appears in both may be significant. To determine the rarity of a particular usage, the corpus of the Google Books N-gram Database can be consulted to confirm not just whether a usage is rare, but exactly how common or rare it is in natural language.[97] For example, it shows that the phrase "stand in line" is more than thirty times as common as the phrase "stand on line."[98] There are other useful corpora that can be consulted in assessing the characteristics displayed in a text, including, *inter alia*, the Corpus of Contemporary American English.[99] Even Wikipedia can function as a corpus, as it is a large collection of natural language written by huge numbers of individuals. In contrast, forensic odontology has no corpus of the distribution of tooth and mouth characteristics to use as a baseline; instead, a court will have to take the word of a forensic odontologist that a particular bitemark is a reflection of some rare tooth or mouth configuration. Without an objectively collected corpus, fact-finders should not give such opinions much credence.

PCAST's report, mentioned above, noted that the only effective way to address the problem of invalid forensic testimony is to undertake empirical validity testing to determine whether, to what degree, and under what circumstances feature comparison forensic practices generate reliable

96. Computer programs that tokenize a text can automate analysis of the words in a text, the character types, the syntactic features of a text such as its sentence and phrase structures, and semantic features of a text such as semantic dependencies and synonym choice. Stamatatos, *supra* note 56, at 539.

97. Jean-Baptiste Michel et al., *Quantitative Analysis of Culture Using Millions of Digitized Books*, 33 SCIENCE 176, 176–82 (2011).

98. Support for this contention was drawn for the year 2000 from Google Books N-Gram Viewer. *Google Books N-Gram Viewer*, GOOGLE, https://books.google.com/ngrams (last visited Dec. 6, 2017). The raw numbers are 0.0000172110% and 0.0000005466% of the corpus, respectively; the ratio between the two is approximately 31.5 to 1. This example shows the remarkable accuracy with which baselines in language usage can be established from corpora.

99. Mark Davies, *The 385+ Million Word* Corpus of Contemporary American English *(1990–2008+): Design, Architecture, and Linguistic Insights*, 14 INT'L J. CORPUS LINGUISTICS 159, 159–90 (2009).

conclusions.[100] For pattern-comparison forensic fields, tests should be established to develop (1) error rates for methods, and (2) likelihood ratios for conclusions, using data sets similar to those that would appear in forensic contexts. The administrator of the validation test should not know that correct answer in advance. Such "blind" testing is essential to mitigating the influence of confirmation bias on validity tests.

It is also important to devote some time to discussing the influence of confirmation basis on forensic science practices. You might think that forensic experts would not be susceptible to confirmation bias—after all, they see themselves as performing science-based assessments, and are typically familiar with the scientific method and its vulnerabilities. The research of Itiel Dror and his colleagues suggests, however, that we cannot assume that forensic experts are immune to confirmation bias. Dror conducted two studies of fingerprint analysts to see if their analyses of whether a latent fingerprint matched a reference print were influenced by confirmation bias. In the first test, he asked five highly-experienced analysts to examine fingerprints, after telling them that the fingerprints were from a high profile case in which FBI fingerprint analysts had erroneously identified an Oregon attorney, Brandon Mayfield, as being a match for prints left in the Madrid terrorist bombing of a train.[101] The experts were asked to examine the fingerprints in question, and to assess whether the prints were a match or not. Each analyst was cautioned to ignore that fact that these were the now-known-to-be-incorrectly-matched Mayfield prints and to simply make an objective analysis of the evidence. In reality, however, the fingerprints each analyst was given were not the Mayfield prints; each analyst was actually given prints that he himself had analyzed in an earlier case and in that case unequivocally called a match. Three of the five experts examining the Mayfield prints reported that, in their opinions, the prints definitely did not match. A fourth expert could not decide if the prints matched or not, and only one of the experts agreed with his original assessment—that the prints appeared to be a match. In other words, when primed with information that the prints did not actually match, four of the five analysts changed their minds about evidence that in an earlier case they had no doubt about.

An objection to the Mayfield-related confirmation bias experiment could be that the biasing information used in that case—from a cause célèbre among fingerprint analysts—was unnaturally powerful in causing them to disbelieve their own observations. To address this possibility, Dror

100. PCAST REPORT, *supra* note 29, at 5–6.

101. Itiel E. Dror et al., *Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications*, 156 FORENSIC SCI. INT'L 74 (2006).

conducted a second experiment, this time providing biasing information typical for forensic practitioners to be exposed to while performing an analysis.[102] In this experiment, six certified fingerprint analysts who had passed stringent proficiency testing were selected. Each was given eight sets of latent and reference fingerprints to evaluate in what they were told were new cases. In reality, however, the fingerprints were from earlier cases in which the analysts had determined the prints to be either clear matches or clear non-matches. Half of these purportedly new cases were control cases, with no biasing information given. For the other half of the cases, the analyst was told potentially biasing information, being either that the suspect in the case had already confessed, or that the suspect was known to have been in jail at the time of the crime. Like the Mayfield test, this experiment showed clear evidence of confirmation bias. Two thirds of the analysts given biasing information rendered answers inconsistent with their earlier assessment of the identical prints.

Follow-up research has shown that cognitive confirmation bias affects more than just forensic fingerprint analysts.[103] A survey of more than 400 forensic practitioners revealed that most had only a limited understanding of confirmation bias, few believed that their own assessments could be affected by confirmation bias, and the few willing to consider that they might be subject to confirmation bias often believed that they could overcome any such bias purely through will power.[104] The demonstrated influence of confirmation bias on forensic practitioners—together with wishful thinking on the part of experts about their own immunity to bias—makes it obvious that blind testing of forensic methodologies is imperative if we are to have reliable forensic science.

Validity testing for error rates is one of the factors that the *Daubert* court held to be relevant in determining the reliability of a forensic field.[105] An authorship analyst can develop a hypothesis—say, that different people will consistently use words of different lengths—and then test that hypothesis on experimental data to determine its validity. For language analysis, it is easy to gather data to test a theory and to see whether—and to what extent—the hypothesis holds true. Because linguistic analysis is nondestructive (unlike, say, arson investigative testing), the same samples can be evaluated

102. *See* Itiel E. Dror & David Charlton, *Why Experts Make Errors*, 56 J. FORENSIC IDENTIFICATION 600 (2006).

103. *See* Saul M. Kassin, Itiel E. Dror & Jeff Kukucka, *The Forensic Confirmation Bias: Problems, Perspectives, and Proposed Solutions*, 2 J. APP. RES. IN MEMORY & COGNITION 42 (2013). For a survey showing that this is a global problem, see Jeff Kukucka et al., *Cognitive Bias and Blindness: A Global Survey of Forensic Science Examiners*, 6 J. APPLIED RES. IN MEMORY & COGNITION 452 (2017).

104. *See* Kassin, *supra* note 103.

105. *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 593–94 (1993).

numerous times, either as a replication check or to compare old methods and protocols against proposed improvements. As a result, it is fairly straightforward to develop specific algorithms to perform authorship analyses, and to instantiate these algorithms into functional computer programs. Indeed, such efforts have already begun, and examples of such programs include the Delta Spreadsheet, the Java Authorship Attribution Program (JGAAP), J-Stylo, and R-Stylo, as well as algorithms devised by the various participants in the PAN conferences.[106]

Using programs like these, it is possible for one team to replicate the analysis of a second team and to obtain *exactly* the same findings. This eliminates one source of error in forensic validation studies that is often overlooked.[107] Researchers may accidentally make an error in conducting the protocols of a test; such errors are often difficult to detect because of changes that testing can inflict on the original data. It is impossible to replicate a chemical analysis of a bloodstain, as the blood is typically destroyed in the process of testing. Linguistic data, however, is analyzable without making any changes in the data itself, and thus that data can be used by any researcher to confirm or disconfirm the apparent results of any other researcher.

Finally, the most important normative aspect of stylometric authorship analysis is its disciplinary commitment to objective testing and the validation of proposed feature sets and comparison methods. It has become a common practice within the field for researchers to measure the performance of a proposed new method on a set of reference documents by using repeated analyses of shared documents with associated ground truths. This practice helps to create a publicly shared estimate of the accuracy of both established and emerging methods. This makes it possible to measure progress in the field, as practices with lower error rates for a particular type of text analysis problem come to replace less accurate practices. In contrast, many feature comparison forensic practices currently lack methods for objectively comparing competing practices within their fields, making it impossible to determine whether a new technique or practice is an improvement on existing practices. The norms and practices of forensic stylometric researchers—their commitment to openness and sharing of data

106. David Hoover, *The Delta Spreadsheet*, 20 LITERARY & LINGUISTIC COMPUTING 20 (2005); Patrick Juola, John Sofko & Patrick Brennan, *A Prototype for Authorship Attribution Studies*, 21 LITERARY & LINGUISTIC COMPUTING 169, 169–78 (2006); Maciej Eder, Jan Rybicki, & Mike Kestemont, *Stylometry with R: A Package for Computational Text Analysis*, 8 R J. 107 (2016).

107. *See, e.g.*, Angi M. Christensen et al., *Error and Its Meaning in Forensic Science*, 59 J. FORENSIC SCI. 1231 (2014) (noting a variety of kinds of error in forensic practices, including errors inherent in the methodology, errors in applying the methodology, errors in assessing results, errors in conducting the practice, measurement errors, etc.).

and methods, to validity testing of methodologies, and to the necessity of continuing to improve accuracy and reliability of stylometric assessments—have established forensic authorship attribution practices as a forensic science truly worthy of that name.

VII. CONCLUSION: THE ONLY RULE IS, IT HAS TO WORK[108]

We borrow the tag line for our concluding section from a book by baseball analysts Ben Lindbergh and Sam Miller, who adhere to the "sabermetric" school of baseball research, advocating the use of empirical data to determine the best tactics to use in baseball games. The book recounts the authors' experiences when they had the opportunity to run an independent league baseball team for a season. Thrilled to have the chance to put their analytic theories into practice, they quickly realized that the opportunity came with a heavy responsibility. Theories that didn't work might make the players look bad, or even torpedo their chances of advancement in their baseball careers. Because their choices could impact the futures of real people, Lindbergh and Miller decided to implement only those moves that "had to work".

We, too, believe that the fundamental criterion for the admission of forensic evidence should be this: Is there evidence that the methodology, as applied to the data in this case, works? Put another way: Does the methodology generate valid and reliable results? In the case of feature identification evidence, that means testing the methodology on data similar to what would be expected in a litigation context, where the scientist does not know the actual answer (as in a real dispute), but where the ground truth is known to someone other than the forensic scientist. As we have shown, testing of this kind is the norm in authorship attribution analyses, providing assurance that this forensic science both can and will lead to more accurate fact-finding. Such tests also generate an accurate estimate of the error rate of the target methodologies, serving as a baseline for researchers seeking to improve their methodologies and revealing which are most useful in solving particular kinds of problems. The use of validity measurements can help to improve the administration of justice, both by helping judges make better admissibility decisions, and by helping factfinders weigh evidence appropriately.

When the accuracy of a forensic science is being constantly and visibly developed, practitioners may be assured that the evidence it produces promotes rather than frustrates justice. Of course, there is always the

108. BEN LINDBERGH & SAM MILLER, THE ONLY RULE IS IT HAS TO WORK: OUR WILD EXPERIMENT BUILDING A NEW KIND OF BASEBALL TEAM (2016).

possibility that testing will disconfirm the reliability of a forensic practice; indeed, this happened in the case of forensic metallurgic analysis of bullets. For decades, the FBI trained crime lab technicians to analyze the metallurgic components of bullets found at a crime scene and then to compare the results to analyses of other bullets—say, for example, those in the possession of the defendant. The technicians would then testify that the crime bullets and the defendant's bullets had matching metallurgic characteristics, sometimes even going so far as to claim that the bullets taken from the crime scene and from the defendant came from the same box. In 2004, however, a National Academy of Sciences commission undertook a study of such analyses and concluded that drawing conclusions as to the common origin of bullets was unreliable. Based on these findings, the FBI discontinued the use of bullet metallurgic analysis as a recommended forensic practice.[109] While unfortunate for the affected forensic experts, this is how science is supposed to work, with unreliable practices getting disconfirmed by objective testing, and valid science becoming increasingly accurate and reliable.

Law needs forensic science in order to assist judges and juries in accurate fact-finding. But law also needs a dependable way to distinguish between astronomy and astrology, science and pseudoscience. What is needed in order for forensic sciences to fulfill their promise in helping to achieve justice is a commitment to what Jennifer Mnookin and her colleagues have called a "research culture" in forensic practice.[110] In this article, we have presented an example of the development of a robust research culture in the field of forensic authorship attribution. The recent comprehensive reports on the state of forensic science by the National Academy of Sciences and the President's Council of Advisors on Science and Technology stressed the value that validity testing brings to forensics; we strongly endorse that position with respect to forensic authorship attribution.[111]

Mature sciences are evidence-driven and welcome validity testing as fundamental to scientific progress. Erin Murphy, who has written extensively on forensic DNA evidence, points out that "[G]ood science is a dynamic process, not a test to be taken and then forgotten once passed."[112] The hallmark of healthy forensic science is a commitment to continual

109. D.H. Kaye, *The Current State of Bullet-Lead Evidence*, 46 JURIMETRICS J. 99, 99 (2006).

110. *See* Jennifer L. Mnookin et al., *The Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 758 (2010).

111. As Juola has observed elsewhere: "Forensic stylometry has a long history of successes mixed with the occasional failure. Ultimately, empirical testing . . . will be key to determining what specific techniques can minimize this chance of failure, and hence maximize the usefulness of this kind of analysis." Juola, *supra* note 52.

112. Erin Murphy, *What 'Strengthening Forensic Science Today' Means for Tomorrow: DNA Exceptionalism and the 2009 NAS Report*, 9 L. PROBABILITY & RISK 7, 22 (2009).

improvements in techniques and methodologies. The research culture that has developed in authorship attribution analysis should serve as a model for other forensic practices to emulate. If that happens, the resulting improvements in accuracy and reliability that we have seen and continue to see in authorship analysis may well be replicated in other forensic fields, to the benefit of overall accuracy and fairness in our legal system.