Journal of Law and Policy

Volume 21
Issue 2
SYMPOSIUM:
Authorship Attribution Workshop

Article 12

2013

Being Pragmatic About Forensic Linguistics

Edward K. Cheng, J.D.

BEING PRAGMATIC ABOUT FORENSIC LINGUISTICS

*Edward K. Cheng**

If my late colleague Margaret Berger taught me anything about evidence, it was that the field seldom yields easy answers. After all, law is necessarily a pragmatic discipline, especially when it comes to matters of proof. Courts must make their best decisions given the available evidence. They have neither the luxury of waiting for better, nor the ability to conjure up, evidence (or new technologies) that they wished they had.

Scholars, by contrast, are naturally attracted to the ideal, sometimes like moths to a flame. Ideals reflect the values and commitments of our society, and they provide the goals that inspire and guide research. But when assessing a new field like forensic linguistics as a legal academic, one needs to carefully separate the ideal from the pragmatic. For when it comes to real cases, evidence law can ill afford to allow the perfect to be the enemy of the good.

Bearing this admonition firmly in mind, this article aims to provide some legal context to the Authorship Attribution Workshop ("conference"). In particular, I want to offer some *pragmatic* observations on what courts will likely demand of forensic linguistics experts[1] and tentatively suggest what the field should aspire to in both the short and long run.

---

* Professor of Law, Vanderbilt Law School; Ph.D. Candidate, Department of Statistics, Columbia University. Thanks to Larry Solan for organizing this remarkably interdisciplinary conference and to Dashiell Renaud for research assistance.

[1] While "forensic linguistics" may encompass a broader set of techniques, I will use the term synonymously with the use of linguistic methods for purposes of attributing authorship, the focus of the conference.

I. *DAUBERT*

No discussion of scientific evidence—at least no discussion of scientific evidence in the United States—can begin without referencing *Daubert v. Merrell Dow Pharmaceuticals*.[2] *Daubert* establishes a five-factor test for the admissibility of scientific evidence: i) falsifiability and testing; ii) publication and peer review; iii) error rates; iv) standards; and v) general acceptance.[3] Unfortunately, applying these factors to many of the forensic linguistic methods presented at this conference immediately raises concerns. The methods do not have rigid procedures that have been tested or have known error rates. Excepting the contributions in this issue of the *Journal of Law and Policy*, few have ever been published. And, almost by definition, since forensic linguistics is an emerging field, many techniques lack general acceptance.

The principal issue is not that forensic linguistic methods are junk. Rather, the problem is that forensic linguistic methods often change from one case to another to account for case-specific contours: Malcolm Coulthard's case study involved selecting certain misspellings and word choices made over e-mail,[4] while Tim Grant's study explored the peculiar grammar of text messaging.[5] The result is a "moving target," and while moving targets are not necessarily bad as a theoretical matter, they are a big problem for the *Daubert* test, which envisions standardized, broadly applicable (and broadly applied) techniques.

Does this mismatch spell doom for the field? Will forensic linguists thus inevitably face widespread opposition and exclusion by judges? Emphatically no. As many in the scientific evidence community have long observed, *Daubert* in practice fundamentally differs from *Daubert* in theory. In real life, courts often treat the *Daubert* factors more as incantation than as actual

---

[2] Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993).

[3] *See id*. at 593–94.

[4] Malcolm Coulthard, *On Admissible Linguistic Evidence*, 21 J.L. & POL'Y 441 (2013).

[5] Tim Grant, *TXT 4N6: Method, Consistency, and Distinctiveness in the Analysis of SMS Text Messages*, 21 J.L. & POL'Y 467 (2013).

requirements.[6] What they really impose is an impressionistic type of scrutiny, giving the expert testimony a "hard look" for intellectual rigor, but nothing more.

Courts have gravitated toward hard-look scrutiny not out of laziness or ignorance[7] but out of pragmatism. The *Daubert* case itself arose in the pharmaceutical context, where large datasets, standardized treatments, and statistical studies reign. What the *Daubert* test demands is thus perfectly reasonable in that context. In other contexts, however, useful expertise exists in the absence of such data. For example, like forensic linguists, accident reconstruction experts also customize their analyses based on case specifics. This customization again means little standardization or statistical justification. Yet, courts have regularly admitted reconstruction experts under hard-look review.[8]

The contours of this hard-look test seem to boil down to three somewhat related inquires. First, is the expert overselling the power of his technique? Courts display little patience with expert grandstanding, strongly preferring ones who carefully delineate what their techniques can and cannot do.[9] Second, does the expert provide a rational explanation for how the technique works? *Daubert* is in many ways an emphatic rejection of *ipse dixit* or say-so testimony.[10] Even though jurors lack technical expertise, *Daubert* tasks them with engaged, reasoned, critical decision making. Blind deference to the authority of a well-

---

[6] *Cf.* 5 DAVID L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY § 43:10, at 782 (2012) ("As a result, the Daubert factors have become something akin to incantation in the structural engineering context, rather than a roadmap for rigorous inquiry.").

[7] *But see* Sophia I. Gatowski et al., *Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-*Daubert *World*, 25 LAW & HUM. BEHAV. 433, 454–55 (2001) (suggesting that many state court judges may not fully understand the *Daubert* factors).

[8] *See* FAIGMAN ET AL., *supra* note 6, § 44:10, at 810 ("[C]ourts take a pragmatic view, admitting [accident reconstruction] testimony even when testing is absent or is otherwise imperfect or flawed.").

[9] *See id.* §§ 45:4–7 (discussing flaws in expert economic analyses).

[10] Gen. Elec. Co. v. Joiner, 522 U.S. 136, 146 (1997) ("[N]othing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert.").

credentialed expert simply will not do.[11] Finally, is the expert willing to acknowledge and address criticisms of his technique? Overdefensiveness or blithely ignoring well-founded objections often betrays a certain lack of understanding, another worthy ground for exclusion.[12]

Viewed in this light, there is little surprise that courts have generally permitted the linguists at this conference to testify in court,[13] and this trend will likely continue. At least within this hand-picked subpopulation, the experts do not oversell their wares and carefully circumscribe the conditions under which their methods apply. They provide reasoned explanations, and I suppose the mere fact of their attendance at this conference demonstrates a profound commitment to taking objections seriously.

II. A (LONG-TERM) WISH LIST

As argued above, courts are likely to admit forensic linguistics as it currently stands. But presumably, this conference's focus is not merely this basic doctrinal question. Rather, Larry Solan's vision was to consider what forensic linguistics might become and how the field might best aid the legal system.[14] In this aspirational vein, let us therefore consider

---

[11] *See generally* Ronald J. Allen & Joseph S. Miller, *The Common Law Theory of Experts: Deference or Education?*, 87 NW. U. L. REV. 1131 (1993) (discussing whether the role of experts is to educate the jury or to arrive at conclusions to which a jury defers).

[12] *Cf.* FAIGMAN ET AL., *supra* note 6, § 43:14, at 786–87 (discussing the courts' use of "robustness tests," which test how well an expert addresses alternative theories or contrary evidence, in the structural engineering context).

[13] Perhaps the most striking example is Carole Chaski, who reports having been allowed to testify in a *Frye* state even after noting repeatedly that her method was experimental and still under development, a condition clearly at odds with her methods being "generally accepted"—the sole criterion for admissibility under a *Frye* test. *See* Carole Chaski, *Best Practices and Admissibility of Forensic Author Identification*, 21 J.L. & POL'Y 333, 358 (2013). The suspicion, naturally, is that even in *Frye* jurisdictions, what matters to courts is not the headcount associated with a method but the intellectual rigor of the method as probed by the hard-look test.

[14] Lawrence Solan, *Intuition Versus Algorithm: The Case of Forensic*

a "wish list" of attributes that the law might want from the field. In an ideal world, we would probably like forensic linguistic analysis to have:

- a widely adopted, predefined algorithm (preferably automated);
- a large, random sample of known exemplars (preferably subclassified by topic and genre); and
- a well-understood theoretical underpinning.

These goals are not my brainchild but have been implicit in many comments, criticisms, caveats, and apologies heard throughout this conference. We all seem to wish that forensic linguistics had fewer ad hoc, case-specific methods so that we could have more rigorous testing and known error rates. We wish that we had a larger and more detailed set of training data so that we could be more confident about external validity. And finally, the linguists, although perhaps not the computational ones, would feel more comfortable if the methods and results were better rooted in linguistic theory.

A moment's reflection suggests the loftiness of these goals. Only one forensic method arguably satisfies them all—DNA. DNA has a widely adopted, predefined, largely automated algorithm; a large, random sample of known exemplars; and a well-understood theoretical underpinning. That is not to say that its history and development were without controversy,[15] but that is where matters stand today. No other forensic field can make such claims.

Juxtaposed to DNA, forensic linguistics clearly has a long way to go. Nearly all of the procedures and algorithms presented at this conference involve some degree of ad hoc expert tweaking and customization, particularly those used for short writing samples. The computational procedures that

---

Authorship Attribution, 21 J.L. & POL'Y 551 (2013).

[15] For example, forensic DNA evidence generated two National Academy of Sciences reports in rapid succession. The first, published in 1992, failed to resolve controversies that were later largely put to rest in the second, published in 1996. *See* NATIONAL RESEARCH COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 10–11 (1996) ("[W]e agree with many recommendations of the earlier [report] but disagree with others.").

involve less tweaking ideally require a large, random sample of exemplars that currently does not exist. And in almost all cases, the theoretical underpinning for the results is opaque. For example, participants offered some off-the-cuff rationales for why *n*-grams[16] or the other machine learning methods work, but no one *really* understands what is going on.

These ultimate goals are surely daunting, but we should be encouraged that the leaders in the forensic linguistics community have set their sights correctly on the prize.

III. SHORT-TERM ASPIRATIONS

With the long-term goals set, let us consider what courts might demand from forensic linguistics in the short term. As I mentioned in the introduction, the legal system must be more pragmatic in the short term, so what exactly should it demand? In this context, *Daubert* hard-look review in conjunction with the other evidentiary rules provides a convenient short-term checklist for forensic linguists.

1. The testimony must add value. This requirement is at the heart of the relevance standard established by Rule 401[17] and the "help the trier of fact" standard governing experts under Rule 702.[18] At the very minimum, forensic linguists should be more than highly credentialed window dressing on common sense. They must add substantive value.

This requirement appears easily met, especially when the expert moves beyond obvious identifying features such as misspellings or unusual word choices. For example, techniques exploiting syntactic structure, choice of function words or grammar, or *n*-grams clearly represent ideas beyond the ken of the average (or even sophisticated) juror.

---

[16] An *n*-gram is a sequence of *n* adjacent items—words, phrases, or characters—from a given text, forming the basis for analysis.

[17] FED. R. EVID. 401.

[18] FED. R. EVID. 702(a) ("A witness who is qualified as an expert . . . may testify in the form of an opinion or otherwise if: (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue . . . .").

2. The testimony must enlighten more than distort or confuse. This second requirement has both evidentiary and statistical inspirations. Evidentiarily speaking, Rule 403 requires that the probative value of evidence not be substantially outweighed by its potential for unfair prejudice, confusion of the issues, or misleading the jury.[19] Statistically speaking, George Box's well-known maxim sums up the problem in a nutshell: "[A]ll models are wrong, but some are useful."[20]

On this score, somewhat counterintuitively, the trend toward quantitative and statistical measures may be more worrisome than more traditional, off-the-cuff qualitative methods. To be sure, as Jay Koehler notes, qualitative methods present hazards through loaded and vague words like "match" and "consistent."[21] But jurors are at least more comfortable weighing that kind of evidence, and attorneys educated about these issues can effectively attack them.

Statistical measures of linguistic similarity are another matter. Statistical methods always have underlying assumptions and potential problems, and asking jurors (or even opposing counsel) to ferret out the distortions created by flawed models is unrealistic. Unless the method is so well-trodden and well-accepted that a jury can essentially use its results uncritically, I worry that statistical models in this context may distort more than illuminate.

3. The testimony must be sufficiently transparent to permit reasoned decision making. This third requirement originates from *Daubert's* hard-look test, as well as Rule 702's demand that a conclusion not rest solely on the *ipse dixit* of an expert.[22]

All of the experts at this conference would presumably meet this criterion with ease, since they have all cogently explained and defended their methods. I can envision two instances, however, in which forensic linguistic testimony could run afoul

---

[19] FED. R. EVID. 403.

[20] GEORGE E.P. BOX & NORMAN R. DRAPER, EMPIRICAL MODEL-BUILDING AND RESPONSE SURFACES 424 (1987).

[21] Jonathan J. Koehler, *Linguistic Confusion in Court: Evidence from the Forensic Sciences*, 21 J.L. & POL'Y 515, 534 (2013).

[22] FED. R. EVID. 702; *see also* Gen. Elec. Co. v. Joiner, 522 U.S. 136, 146 (1997).

of this requirement. The first is the purely impressionistic linguist, who relies solely on his or her "training and experience." Lest this example seem like a straw man, let me note that authentication attempts in other fields frequently proceed along these lines. For example, art experts studying the Getty kouros reported feeling an inexplicable revulsion upon first seeing the statue, and these gut feelings often provided a foundation for their assessment that the statue was a fake.[23] Such intuitions are surely not nonsense, and arguably the legal system should prefer an art expert's opinion over the average juror's, but *Daubert* makes clear that *ipse dixit*, "blink"-type testimony does not make the cut.[24]

The second potentially problematic instance is where a machine-learning algorithm arrives at an empirically successful identification rule (i.e., high accuracy), but researchers have little idea why it works as a matter of substantive theory.[25] With its emphasis on predictive accuracy over interpretability, machine learning tends toward such black boxes, and while I personally sympathize with the approach, the legal system with its emphasis on reasoned decision making typically does not.

4. The method must have some proven empirical validity. This final requirement is based again on the text of Rule 702[26] but may be the most difficult short-term aspiration for the field. The *sine qua non* of empirical validity is testing. For some of the data-intensive, quantitative methods presented at this conference, a focus on testing is practically inherent. But

---

[23] Georgios Dontas, *The Getty Kouros: A Look at Its Artistic Defects and Incongruities*, *in* THE GETTY KOUROS COLLOQUIUM 37, 37 (Angeliki Kokkou ed., Alex Doumas trans., 1993) ("In the controversy regarding the authenticity of the Getty kouros a factor that must be taken into account is, in my opinion, the unfavourable feeling it arouses at the very first glance."); *see also* MALCOLM GLADWELL, BLINK 3–8 (2005) (discussing the Getty kouros).

[24] Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 588–89 (1993).

[25] *See generally* Leo Breiman, *Statistical Modeling: The Two Cultures*, 16 STAT. SCI. 199 (2001) (discussing the two cultures of statistics: one focused on explanation, and the other on prediction).

[26] FED. R. EVID. 702 ("A witness who is qualified as an expert . . . may testify in the form of an opinion or otherwise if: . . . (c) the testimony is the product of reliable principles and methods . . . .").

methods such as those proposed by Coulthard[27] or Grant,[28] which are more qualitative, subjective, or case-specific, will require experts to embrace proficiency testing and out-of-sample testing more affirmatively.

For qualitative linguistic experts, courts should demand proficiency testing—tests of ability involving known problems given under blinded conditions.[29] Such testing is undoubtedly no fun for the experts involved. The experts open themselves up to attack if the testing turns out badly, and the risk of endangering a lucrative line of business creates substantial disincentives to participate. Experts will thus require judicial prodding, for without such information about accuracy rates, jurors cannot assess the probative value of an expert's conclusions.

For case-customized models, any reported accuracy rates must be out-of-sample accuracy rates. Constructing models that merely fit the data on hand is one thing; successfully predicting future data is an entirely different matter. Tailoring methods or models to a specific case is a time-honored recipe for creating overfitted models, which explain the current dataset well but handle future datasets poorly. To get proper accuracy rates, researchers must divide their dataset into training and testing sets. Models should be developed only with the training set, and validation should be done only with the separate testing set. Some of the conference papers used out-of-sample testing, while others either did not or were unclear.[30]

Finally, part and parcel of testing is the establishment of standardized procedures. As the forensic linguistics field matures, it will have to sacrifice some of its flexibility for

---

[27] Coulthard, *supra* note 4.

[28] Grant, *supra* note 5.

[29] Proficiency testing has been proposed as the solution to *Daubert* in other contexts involving subjective, expert-dependent determinations, such as fingerprints. *E.g.*, Jennifer L. Mnookin, *The Courts, the NAS, and the Future of Forensic Science*, 75 BROOK. L. REV. 1209, 1217–33 (2009).

[30] *E.g.*, Shlomo Argamon & Moshe Koppel, *A Systemic Functional Approach to Automated Authorship Analysis*, 21 J.L. & POL'Y 299, 313 tbl.1 (2013) (uses cross-validation); Chaski, *supra* note 13, at 353 tbl.3 (uses cross-validation); Coulthard, *supra* note 4 (does not use cross validation); Grant, *supra* note 5 (does not use cross-validation).

standardization, both across cases and ultimately across experts. Standardization of the feature set used in forensic linguistic analysis is imperative if we are to have established error rates. It is also the only way to avoid confirmation bias. Without a predefined algorithm, an expert runs the significant risk of preferencing aspects that confirm her initial hypothesis over those that disprove it.[31]

Going forward, the challenge for forensic linguists will be to develop a method that relies less on the expertise of the individual linguist—at least on an everyday basis. The heavy lifting in developing an authorship attribution technique should occur in the lab, long before it is applied in a legal case. By the time it is applied for legal consequence, the application of the method should be largely mechanical.

CONCLUSION

Ours is an extremely exciting time for forensic linguistics. The field faces profound challenges in its attempt to meet the ideals and goals set by *Daubert*, and much work remains to be done. Yet, with so many motivated and intellectually engaged scholars and researchers, we can be very hopeful that progress will be steadily made.

More broadly, as a legal observer, I am curious to see how the field of forensic linguistics ultimately develops. Unlike most forensic fields, which arose long before the invention of DNA typing and the decision in *Daubert*, forensic linguistics will blossom within a modern scientific evidence framework. It will thus provide a unique opportunity to observe how the various actors and modern incentives interact. More importantly, it will help evidence scholars determine whether all the trouble collectively known as *Daubert* is really worth the candle.

---

[31] In this context, I am reminded of the modus operandi arguments made by the prosecution in *United States v. Trenkler*, 61 F.3d 45 (1st Cir. 1995), a case involving the purported "signature" of a bomber. The prosecution pointed to several common bomb parts in its argument that two bombs were constructed by the defendant. The dissent rightly wondered why one should emphasize the similarities between the two bombs rather than several significant dissimilarities. *Id.* at 64 (Torruella, J., dissenting).