BEST PRACTICES AND ADMISSIBILITY OF FORENSIC AUTHOR IDENTIFICATION

*Carole E. Chaski**

I. INTRODUCTION

Forensic linguistics provides answers to four categories of inquiry in investigative and legal settings: (i) identification of author, language, or speaker; (ii) intertextuality, or the relationship between texts; (iii) text-typing or classification of text types such as threats, suicide notes, or predatory chat; and (iv) linguistic profiling to assess the author's dialect, native language, age, gender, and educational level. This article discusses author identification in relation to linguistics, research, and admissibility as evidence in U.S. courts.

Federal and states courts in the United States have undertaken three main approaches in determining whether to admit, partially admit, or exclude forensic authorship identification evidence. These three approaches are forensic computational linguistics, forensic stylistics, and stylometric computing. Each has a distinct origin. Forensic computational linguistics developed out of linguistic theory and computational linguistics.[1] Forensic stylistics developed out of traditional forensic handwriting identification.[2] The stylometric computing approach developed out of both literary authorship identification and machine-learning-based text classification.[3]

This article focuses upon the forensic computational linguistic approach and contrasts this approach to the forensic stylistics and stylometric computing approaches. In Section II, best practices for

* Institute for Linguistic Evidence; ALIAS Technology LLC, Georgetown, DE; Ph.D., Brown University.

[1] *See* Carole E. Chaski, *Who Wrote It? Steps Toward a Science of Authorship Identification*, NAT'L INST. JUST. J., Sept. 1997, at 15, 18 [hereinafter Chaski, *Who Wrote It?*].

[2] *See* GERALD R. MCMENAMIN, FORENSIC STYLISTICS 45–46 (1993).

[3] *See* Moshe Koppel & Jonathan Schler, *Exploiting Stylistic Idiosyncrasies for Authorship Attribution*, PROC. IJCAI'03 WORKSHOP ON COMPUTATIONAL APPROACHES TO STYLE ANALYSIS & SYNTHESIS, 2003.

forensic linguistics are presented. The best practices provide an evaluative framework for the forensic computational linguistics approach, discussed in Section III; the forensic stylistics approach, discussed in Section IV; and the stylometric computing approach, discussed in Section V. In each section, a discussion of admissibility is included, since best practices should guide both judicial reasoning as well as scientific practice.

II. Best Practices for Forensic Linguistics

Best practices in forensic linguistics are essential to propel the field of authorship identification from an academic or law enforcement sideline consultancy to a real forensic science that is useful to the judicial system. Best practices include factors from both the legal standards for evidence, so as to be useful and address admissibility concerns, and scientific standards for research, so as to be reliable, replicable, and respectable.

Scientifically respectable and judicially acceptable methods for author identification should be:

a. developed independent of any litigation;
b. tested for accuracy outside of any litigation;
c. tested for accuracy on "ground truth" data;
d. able to work reliably on "forensically feasible" data;
e. tested for known limits correlated to specific accuracy levels;
f. tested for any errors of individual testing techniques that could cause accumulated error when combined with other techniques;
g. replicable;
h. related to a specific expertise and academic training;
i. related to standard ("generally accepted") techniques within the specific expertise and academic training; and
j. related to uses outside of any litigation in industries or fieldwork in the specific expertise.

By implementing these best practices, forensic computational linguistics is oriented primarily toward research- and empirically-driven protocols rather than expert-witnessing. In this way, forensic computational linguistics is a "normal science" subfield of computational linguistics and linguistic theory.[4] Accordingly,

[4] Jennifer L. Mnookin et al., *The Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 725 n.75 (2011). As an example of a

forensic computational linguistics belongs to a thriving community of academic and industry linguists with educational and industrial standards. These best practices go far toward "solving the 'hired gun' problem" that plagues American courts and universities—when academicians do not conduct research at all or research congruent with best practices but make themselves available as expert witnesses.[5]

A. Litigation Independence

The implementation of these best practices involves litigation-independent development and testing of any method on a ground-truth dataset that contains forensically feasible data.[6] The researcher-forensic linguist runs experiments to test how well a method works outside of any litigation. The results are simply what they are, not favoring one side or the other of a legal dispute. Such a testing environment frees the researcher from confirmation bias because the results are simply what they are and enable the researcher to design the next set of experiments, as is usual in

research culture in forensic linguistics, the Institute for Linguistic Evidence, founded in 1998 through funding from the U.S. Department of Justice's National Institute of Justice, is the first research organization devoted to validation testing for methods related to linguistic evidence. *See* INST. FOR LINGUISTIC EVIDENCE, http://www.linguisticevidence.org (last visited Apr. 18, 2013). ILE has embraced the forensic computational linguistic paradigm from its inception and over the years has averaged about five research associates working on an average of four research projects per year. *See id.* Academicians had been functioning as expert witnesses in forensic linguistics since the 1980's. Professor Roger Shuy of Georgetown University was one of the earliest forensic linguistic experts and has described his cases prolifically, but has not sustained a research agenda in the field. Professor Gerald R. McMenamin, another early expert witness in forensic linguistics, has provided both case reports and descriptions of his method, but no testing of the method for error rate. Ironically, the "research culture" that Mnookin et al. fairly state as lacking in forensic science and crime labs is just as lacking for forensic linguistics in the halls of academe. *See* Mnookin et al., *supra*, at 765.

[5] The plague of "hired Guns" or "whores of the court" in the U.S. judicial system has been amply documented in PETER W. HUBER, GALILEO'S REVENGE: JUNK SCIENCE IN THE COURTROOM (1993); *see also* MARCIA ANGELL, SCIENCE ON TRIAL (1997); MARGARET A. HAGEN, WHORE OF THE COURT: THE FRAUD OF PSYCHIATRIC TESTIMONY AND THE RAPE OF AMERICAN JUSTICE (1997).

[6] Carole E. Chaski, *Author Identification in the Forensic Setting*, *in* THE OXFORD HANDBOOK OF LANGUAGE AND LAW 494, 494–99 (2011) [hereinafter Chaski, *Author Identification*].

normal science.

B. Ground-Truth Data

For the testing to be meaningful, the experiments must be run on ground-truth data.[7] A ground-truth dataset contains known, verified examples with features relevant to the experiments being run.[8] For author identification, a ground-truth dataset typically contains text samples for which the authorship is known and verified.[9] For writer identification, a ground-truth dataset typically contains writing samples for which the hand writer is known and verified.[10] For linguistic profiling, a ground-truth dataset typically contains linguistic examples for which the demographics of each author/speaker are known and verified.

It is impossible to calculate a trustworthy accuracy rate if the researcher does not use ground-truth data. Determining a method's accuracy requires comparing the method's results to the correct answers. Correct answers can only arise from ground-truth data, where the dataset is known and verified. If the researcher is using a dataset with 100 texts but an unknown number of authors, he will never know, with complete certainty, how many of those 100 texts his method correctly assigned to the actual author.[11] If the researcher is using a dataset containing 10,000 authors with demographic features, but the researcher has not verified those demographic features, he will never accurately know how many of those 10,000 authors his method assigned correctly to a gender, age group, or educational level.[12] Essentially, working without ground-truth data is a sophisticated form of guessing: it may look scientific, but it is not real science.

C. Forensically Feasible Data

For the methods to work reliably in actual cases, ground-truth

[7] *Id.*

[8] *Id.*

[9] *Id.*

[10] Carole E. Chaski & Mark A. Walch, *Validation Testing for FLASH ID on the Chaski Writer Sample Database*, PROC. AM. ACAD. FORENSIC SCI. ANN. MEETING, 2009.

[11] Chaski, *Author Identification*, *supra* note 6, at 494.

[12] *Id.*

data must be forensically feasible, i.e., the same kind of data that is obtained in actual cases. In actual cases, writing exemplars are messy, ungrammatical, unedited, cross-genre, cross-register, and sparse because people write naturally, across a range of genres and registers. Accordingly, a forensically feasible dataset will contain business letters, love letters, angry rants, narratives, and essays so that the same author can be examined writing in different genres and registers. Each genre contributes something different to the dataset. For instance, business letters contain more formal word choice and more conventional spelling and punctuation patterns than personal e-mails, love letters, or angry blog posts. Even the writing medium—handwriting, typewriting, or computer keyboarding—can cause intra-author differences such that lexical, spelling, grammar, or punctuation patterns that occur in one medium typically do not occur in another.[13] In case data, the writing exemplars are typically not edited to any conventional, newspaper, academic, or industrial standards. If the researcher is not using a forensically feasible dataset to test his method, he might be misled into thinking that his method—built to assign clean, grammatical, edited business letters, newspaper articles, or novels—will work accurately on messy, ungrammatical, forensically significant texts. Essentially, building a method without testing it on forensically feasible data simply overgeneralizes a method's ability: it may look scientific because there are some validation tests to refer to, but the validation test results do not prove that the method can work on the data in the case or any forensically feasible data.

Research that focuses on literary classics or edited newspaper articles may develop accurate methods, but these methods must be tested on forensically feasible data before they are borrowed across-the-board for forensic authorship identification. In most cases, literary methods fail to work on forensic data simply

[13] A nice example of how writing media can affect spelling comes from the *Van Wyk* case. *See infra* Part III.D. The contraction of [do not] occurred in two ways: in handwritten documents as [don't] and in typed documents as [don;t]. Typewriter and computer keyboards are different in the placement of the semicolon and apostrophe. The typewriter keyboard requires a shift to get the apostrophe, while a computer keyboard does not. The typist did not use the shift key, producing a typical typing error for novices, while the handwriter never made that kind of mechanical error. The context of this difference was not noted in the forensic stylistics report by Agent Fitzgerald; instead he argued that [don;t] was a unique stylemarker.

because the literacy methods require far longer texts than the forensic case affords. Brevity is a fact of life inherent in forensic authorship identification that cannot be avoided or helped by research that focuses on texts that contains thousands, tens of thousands, or hundreds of thousands of words. Again, using methods that work well on literary texts or newspaper text banks, without independently testing the methods on forensically feasible data, may appear to be scientific because there is published literature in humanities computing to refer to about authorship identification in nonforensic settings,[14] but using such methods is akin to using a screwdriver on a nail—and an unvalidated screwdriver at that.

D. Empirically Established Protocol

In the research environment, the continual testing of a method of forensically feasible, ground-truth data empirically establishes the protocol for using the method in casework. First, a level of accuracy can be set: for instance, the method won't be used forensically until it reaches a certain accuracy level, such as eighty percent, ninety percent, or ninety-five percent. Second, the experiments are designed to control for variables such as the quantity of data, required number of authors, required number and types of linguistic features, and the required number and types of individual testing techniques that are combined in the method.

For the quantity of data, an important issue to resolve is the minimum number of words, sentences, or texts required for the method to obtain a certain level of accuracy.[15] For the number of authors, a method may require a minimum of two, five, or twenty-five suspects to obtain a certain level of accuracy. As in other pattern recognition techniques in forensic science, the number and type of features required for identification or elimination is established empirically by controlling the variable in a series of validation tests related to specific accuracy rates.[16] If fifteen

[14] For instance, the Association for Computers and the Humanities publishes *Literary and Linguistic Computing*, a journal where authorship issues in literature, religion and other nonforensic settings are regularly discussed.

[15] *See* Chaski, *Author Identification*, *supra* note 6, at 503.

[16] Carole E. Chaski, *Empirical Evaluations of Language-Based Author Identification Techniques*, 8 INT'L J. SPEECH LANGUAGE & L. 1 (2001) [hereinafter Chaski, *Empirical Evaluations*]; Carole E. Chaski, *Who's at the*

linguistic features from syntactic analysis yield eighty-two percent accuracy, the next experiment will test sixteen, seventeen and so forth until the desired accuracy level is achieved. Those experiments empirically establish the number and type of features required for the method to obtain a specific accuracy level. Likewise, if measurement or feature selection techniques can be combined in a method (combining syntax with other linguistic features or combining measurement based on word overlap with measurement based on *n*-grams), then experimental tests must be run to determine which techniques and how many techniques must be combined to reach a specific level of accuracy.

By working independently of any litigation and running experiments that control for different variables in how the method can be implemented, the researcher forensic linguist empirically establishes a protocol for each tested method. The protocol then becomes the guidelines for actually using the method in real casework. There will be cases where the tested methods cannot be used because data requirements cannot be met (i.e., a decedent cannot provide more writing samples), and there will be times when the tested methods can be used but only with the caveat that the data requirements for the most robust results are not met fully but are close to being satisfied (i.e., the decedent's writing samples are close to the required number). These types of situations should encourage additional research and not lead to abandonment of the research paradigm. In fact, the empirically established protocol prevents the researcher forensic linguist from becoming a "hired gun" who merely runs a method in whatever way to get the "desired result," rather than in accord with an empirically established protocol that provides a specific level of accuracy outside of litigation.

Note that "having worked a lot of cases" is not at all a substitute for empirically establishing a protocol. It simply means that a person has been hired a lot. The researcher forensic linguist has run a lot of experiments independent of litigation—a state that is far more valuable to developing forensic linguistics into a real

Keyboard? Authorship Attribution in Digital Evidence Investigations, INT'L J. DIGITAL EVIDENCE, Spring 2005, at 1 [hereinafter Chaski, *Who's at the Keyboard?*]; Carole E. Chaski, Presentation at the Eight Biennial Conference on Forensic Linguistics/Language and Law: Empirically Testing the Uniqueness of Aggregated Stylemarkers (July 14, 2007) [hereinafter Chaski, *Empirically Testing*].

and reliable science than a client list.

E. Controlling Cumulative Error

Most methods for forensic author identification require some tools for measurement or feature selection.[17] These tools can produce errors in and of themselves; thus, an accuracy rate can be seriously affected by a series of accumulating errors in measurement or selection. For instance, off-the-shelf parsers developed in academia get very high accuracies for part-of-speech tagging on clean, edited data such as newspaper articles and novels. But these same off-the-shelf parsers often fail miserably on ungrammatical data. The problem of parsing ill-formed input or ungrammatical sentences was first discussed over thirty years ago,[18] and it has not been fully solved.[19] If the method uses an off-the-shelf parser and does not involve checking the parser results and correcting any errors of part-of-speech tagging or phrase chunking, then those errors pass through to the next step of the method. Another set of errors that can be created by software is the common practice of "preprocessing" texts to rid it of extra spaces, or to correct spellings, or insert punctuation. All of these preprocessing maneuvers actually change the original data and could remove some features that are actually useful for author identification. This kind of data handling is not scientifically acceptable even if it makes software run easily, and it undermines the accuracy of any methods that use the "preprocessed" data.

Another example is the interpretation of handwritten symbols: if a stroke is interpreted as an errant apostrophe but it is actually a low comma, this error of interpretation must be corrected, lest a later classification rely on the misinterpretation. As such errors accumulate, the linguistic analysis becomes less and less accurate, so that neither the method's accuracy rate nor the final decision assigning texts to authors can be trusted.

[17] *See* Chaski, *Author Identification*, *supra* note 6, at 491–93.

[18] *See* K. Jensen et al., *Parse Fitting and Prose Fixing: Getting a Hold on Ill-Formedness*, 9 AM. J. COMPUTATIONAL LINGUISTICS 147 (1983); Ralph M. Weischedel & John E. Black, *Responding Intelligently to Unparsable Inputs*, 6 AM. J. COMPUTATIONAL LINGUISTICS 97 (1980); Ralph M. Weischedel & Norman K. Sondheimer, *Meta-Rules as a Basis for Processing Ill-Formed Input*, 9 AM. J. COMPUTATIONAL LINGUISTICS 161 (1983).

[19] *See* Jennifer Foster & Carl Vogel, *Parsing Ill-Formed Text Using an Error Grammar*, 21 ARTIFICIAL INTELLIGENCE REV. 269 (2004).

F. Replicability

The protocol developed through repeated validation testing must be repeatable by others who use it. Methods within the protocol also must be repeatable through the implementation in computer software or the strict operationalization of terms and procedures. Implementing a method in computer software is a sure way of providing objectivity and maintaining consistency. Systems can be designed so that each user can tweak parameters, thereby changing the algorithm. However, these tweaks might not be visible later. Accordingly, such systems do not maintain consistency in running a method, and the fact that a method is implemented in software does not necessarily guarantee that it is completely replicable.

G. The Method's Relationship to Academic and Industrial Uses

Finally, the research environment should be related to academia and/or industry by the sharing of knowledge, techniques, methods, or software. The researcher forensic linguist is part of a larger community of computational linguists, psycholinguists, corpus linguists, theoretical linguists, and computer scientists where forensic applications are just one application of common techniques, methods, and software put together in novel ways. For instance, text classification techniques were originally designed as part of summarization schemes but later became useful for finding plagiarism and duplicates within large electronic collections, just as DNA testing was originally used for paternity before it was applied forensically.

Forensic author identification methods should relate, in some recognizable way, to a theory of language, since the method is seeking to identify authorship based on language (rather than handwriting, ink, or IP address). Linguistics obviously offers the fullest theories of language, with the generative theory being the best developed. The generative theory of language includes Chomsky's original transformational-generative grammar,[20] now known as Minimalism,[21] as well as its offshoots such as Lexical-Functional Grammar;[22] Generalized Phrase Structure Grammar;[23]

[20] *See* NOAM CHOMSKY, ASPECTS OF THE THEORY OF SYNTAX (1965).

[21] *See* NOAM CHOMSKY, THE MINIMALIST PROGRAM (1995).

[22] *See* JOAN BRESNAN, LEXICAL FUNCTIONAL SYNTAX (2001).

Head Driven Phrase Structure Grammar;[24] and Construction Grammar.[25] What has been especially impressive about the generative theory of language is its ability to make predictions about linguistic structure, linguistic functions, and the psychological reality of linguistic structure. Other theories, such as Tagmemics[26] or Systemic Functional Grammar,[27] have remained primarily descriptive or taxonomic rather than predictive.

Prescriptive grammar—or school grammar—is taught in schools to indoctrinate students with the prestige or most socially desirable dialect and especially how to "use words correctly." It teaches how a native speaker *should* speak rather than how a native speaker *actually* speaks. Prescriptive grammar is neither descriptive nor predictive, as it is not a scientific theory of language but is the standard approach to language for literary analysis and for anyone who has not studied linguistics. Prescriptive grammar is attractive to judges who typically write and speak a prestige dialect congruent with prescriptive grammar. However, research has demonstrated that prescriptive grammar is not an adequate theory of language for authorship identification.[28]

Differences in academic training make the paradigm of experimental validation testing for forensic authorship identification more or less difficult to accept. Training in literary criticism does not focus on empirical methods, while pure computer science can bypass courses in experimental design. However, in most branches of linguistics, empirical work is mandatory. Psycholinguists design and run experiments testing the theoretical constructs posited by linguistic theory (usually from a

[23] *See* GERALD GAZDAR ET AL., GENERALIZED PHRASE STRUCTURE GRAMMAR (1985).

[24] *See* CARL POLLARD & IVAN A. SAG, HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR (1994).

[25] *See* THOMAS HOFFMAN & GRAEME TROUSDALE, THE OXFORD HANDBOOK OF CONSTRUCTION GRAMMAR (2013).

[26] *See* KENNETH L. PIKE, LANGUAGE IN RELATION TO A UNIFIED THEORY OF THE STRUCTURE OF HUMAN BEHAVIOR (1967); KENNETH L. PIKE, LINGUISTIC CONCEPTS: AN INTRODUCTION TO TAGMEMICS (1982).

[27] M.A.K. HALLIDAY & CHRISTIAN M.I.M. MATTHIESSEN, AN INTRODUCTION TO FUNCTIONAL GRAMMAR (3d ed. 2004).

[28] *See* Michael Brennan & Rachel Greenstadt, *Practical Attacks Against Authorship Recognition Techniques*, PROC. TWENTY-FIRST CONF. ON INNOVATIVE APPLICATIONS ARTIFICIAL INTELLIGENCE (IAAI), 2009, at 60; Chaski, *Empirical Evaluations*, *supra* note 16; Koppel & Schler, *supra* note 3.

generative theory), focusing on the cognition and memory required to produce and process human language. The validation testing described earlier is second nature to someone trained in psycholinguistics (including child language acquisition, psychology of literacy, and second language acquisition).

Even if the forensic linguist relates to the small community of sociolinguists, the methods that the forensic linguist develops should be recognizable as sociolinguistics. Historically, sociolinguistics introduced a quantitative approach to midcentury American linguistics and relied heavily on empirical data collection, phonetic measurements, and experimental research designs.[29] Therefore, when a forensic linguist asserts that his academic training is in sociolinguistics, but his method is neither quantitative, nor tested on ground-truth data, nor validated by experiments, the disconnect between the forensic activity and the academic world is startling to linguists, if invisible to attorneys or judges.

III. The Forensic Computational Linguistics Approach to Author Identification

Work in forensic computational linguistics began in the mid-1990s, with funding from the National Institute of Justice.[30] By the late 1990s, I had developed a method now known as SynAID (Syntactic Author Identification) within ALIAS (Automated Linguistic Identification and Assessment System).[31] This research has played a role in adjudicated cases in 1998, 2001, and 2008, discussed later.

Litigation-independent validation testing on forensically

[29] Labov is considered the originator of sociolinguistics; his work is characterized by quantitative, statistical analysis of naturally collected or elicited linguistic behavior. *See generally* William Labov, The Social Stratification of English in New York City (2d ed. 2006).

[30] In 1995, I received a grant to validate linguistic methods for determining authorship, Grant ID 1995-IJ-CX-0012, Visiting Fellowship, Linguistics Methods for Determining Authorship.

[31] *See* Chaski, *Empirical Evaluations*, *supra* note 16; Chaski, *Empirically Testing*, *supra* note 16; Carole E. Chaski, Recent Validation Results for the Syntactic Analysis Method for Author Identification, International Conference on Language and Law (2004) [hereinafter Chaski, *Syntactic Analysis Method Identification*]; Chaski, *Who Wrote It?*, *supra* note 1; Chaski, *Who's at the Keyboard?*, *supra* note 16.

feasible ground-truth data is a core feature of the forensic computational linguistics approach. Implementation in software that is responsive to messy data is central to the forensic computational linguistics approach for both replicability and error control. Linguistic theory plays a central role in the forensic computational linguistics approach. These features distinguish the forensic computational linguistics approach in sometimes obvious, sometimes subtle ways from forensic stylistics and stylometric computing.

A. Linguistic Theory Does Matter

In linguistic theory, language is divided into levels for analytical purposes.[32] These levels are sound, word, and word combinations.[33] These levels, respectively, are analyzed in phonetics and phonology; morphology and the lexicon; syntax; semantics and pragmatics; and prosody.[34] These levels have different salience or prominence in processing and especially imitation of language. For instance, children acquire sounds and prosody before they acquire words.[35] Syntactic form—or the actual ordering and combination of words—is least salient and consequently least easy to imitate. There was a great deal of research in psycholinguistics starting in the 1960s, none of which has been refuted, about the way we remember the meaning of a statement while we forget how the statement was actually said.[36] In fact, in normal linguistic processing it appears that loss of syntactic

[32] This division of language into analytical levels is commonplace in standard textbooks in linguistics. *See e.g.*, RICHARD AKMAJIAN ET AL., LINGUISTICS: AN INTRODUCTION TO LANGUAGE AND COMMUNICATION (6th ed. 2001); EDWARD FINEGAN, LANGUAGE: ITS STRUCTURE AND USE (6th ed. 2012); VICTORIA FROMKIN ET AL., AN INTRODUCTION TO LANGUAGE (10th ed. 2013).

[33] *See* AKMAJIAN ET AL., *supra* note 32; FINEGAN, *supra* note 32; FROMKIN ET AL., *supra* note 32.

[34] *See* AKMAJIAN ET AL., *supra* note 32; FINEGAN, *supra* note 32; FROMKIN ET AL., *supra* note 32.

[35] S. Katz-Gershon, Word Extraction in Infant and Adult Directed Speech: Does Dialect Matter? (2007) (unpublished Ph.D. dissertation, Wayne State Univ.) (on file with author).

[36] Philip N. Johnson-Laird & Rosemary Stevenson, *Memory for Syntax*, 227 NATURE 412 (1970) (citing Jacqueline S. Sachs, *Recognition Memory for Syntactic and Semantic Aspects of Connected Discourse*, 2 PERCEPTION & PSYCHOPHYSICS 437, 437 (1967)).

structure occurs within milliseconds,[37] even in writing tasks.[38] Nonetheless, even though we do not remember the word order for long, syntactic structures are very real, albeit fragile and abstract. Again, a great deal of research in psycholinguistics and linguistic theory (starting with Fodor and Bever[39]) demonstrates the reality of syntactic structures, especially the edges of structures, like the beginnings and endings of noun phrases or clauses, because the edges are where most informative morphosyntactic elements appear, and also where the phrasal head—the dominant function—is placed. Therefore, the forensic computational linguistic approach focuses primarily on syntax because syntax would be more difficult to imitate than lexical choices or spelling and punctuation (the graphic correlate of phonetics and prosody).

Another aspect of linguistic theory essential to author identification is the theory of markedness.[40] In many human characteristics, there is an asymmetry in function of symmetrical design. Handedness and footedness are the obvious examples of this asymmetry, but the brain also has this kind of duality.[41] Language is permeated from phonetics through pragmatics with asymmetric oppositions, a fact that was first realized and articulated by the Prague School in the 1940s and then adopted within generative linguistics in phonology[42] and in syntax.[43] Markedness explains why some noun phrase structures are harder to process, produce, or find in high frequency while other nouns phrase structures are a dime a dozen, even in child language. [44] A

[37] *Id.*

[38] *See* Holly P. Branigan et al., *Syntactic Priming in Written Production: Evidence for Rapid Decay*, 6 PSYCHONOMIC BULL. & REV. 635, 635–40 (1999).

[39] Jerry A. Fodor & Thomas G. Bever, *The Psychological Reality of Linguistic Segments*, 4 J. VERBAL LEARNING & VERBAL BEHAV. 414, 414–20 (1965).

[40] For an overview of markedness theory in linguistics, see generally EDWIN L. BATTISTELLA, MARKEDNESS: THE EVALUATIVE STRUCTURE OF LANGUAGE (1990).

[41] Kenneth Hugdahl, *Symmetry and Asymmetry in the Human Brain*, 13 EUR. REV. 119, 119–33 (2005).

[42] *See* NOAM CHOMSKY & MORISS HALLE, THE SOUND PATTERN OF ENGLISH (1968).

[43] Judith Aissen, *Markedness and Subject Choice in Optimality Theory*, 17 NAT. LANGUAGE & LINGUISTIC THEORY 673, 673–711 (1999); *see also* GERALD GAZDAR ET AL., GENERALIZED PHRASE STRUCTURE GRAMMAR (1985); CARL POLLARD & IAN A. SAG, HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR (1994).

[44] *See* BATTISTELLA, *supra* note 40.

noun phrase "*the tippy cup with your name on it that we found under the car seat yesterday*" is marked; the noun phrase "*your tippy cup*" is unmarked. Marked noun structures occur later in language acquisition and even in adult language are less frequent than unmarked noun structures.

In phonetics, normalization is the process of speaker recognition by which we come to recognize specific phonetic features in an individual's voice—features that are consistent with the person but also different from someone else.[45] If recognition is possible at the phonetic level—and everyone has had the experience of recognizing a person by voice over the telephone—it is a testable hypothesis that a similar recognizability would be possible at the syntactic level. The issue is to find, again borrowing from phonetics, some invariant signal among the variation and noise (in an information theoretic sense).[46] Or borrowing from statistical terminology, what syntactic patterns would be distinctive enough among the potential note writers to differentiate intrawriter variation from interwriter variation?

Language is a conventional behavior where for the sake of mutual understanding we share the same code. In information theoretic terms, each of us is both sender and receiver. This is how we manage to finish each other's sentences: we are using the same code we share with another person in our linguistic circle. So the notion that individual language is unique, or that each of us has a unique linguistic behavior, is an idea that linguistics as a discipline denies by the very definition of language as a conventional behavior and shared code.

Even though linguistic behavior cannot be literally unique, it can and does show variation. By definition, dialect is the name for group-level linguistic behavior, where subgroups within the language can be determined. At the individual level, linguistics has posited the notion of idiolect, or a variation of language at the individual level.[47] Clearly, idiolect cannot be a unique language, or, again, the unique language would have a speaker of one, but variations at the individual level might still be discoverable.

[45] For an overview of speaker recognition, see Homayoon Beigi, *Speaker Recognition*, *in* BIOMETRICS 3, 3–29 (2011), *available at* http://www.intechopen.com/books/biometrics/speaker-recognition.

[46] *Cf.* CLAUDE E. SHANNON & WARREN WEAVER, THE MATHEMATICAL THEORY OF COMMUNICATION (1971).

[47] *See, e.g.*, FROMKIN ET AL., *supra* note 32.

Idiolect was first posited at the phonetics level. The biological substrate of phonetic articulation certainly makes phonetic individual differences feasible.[48] Idiolect later became a useful theoretical term in recognizing syntactic variation between syntacticians. There is still no empirical method for demonstrating that each person has his or her own idiolectal variation that is uniquely identifiable, but author identification merely has to recognize intrawriter vs. interwriter variation strong enough to differentiate authors from each other and cluster documents by author.[49]

Finally, due to the brevity of the texts, a realistic forensic author identification method needs a way of measuring the texts to get as much information as possible out of them. Counting syntactic structure rather than words yields a higher count and makes statistical analysis possible. If a method only counts the words, the result is a long list of words with frequencies that are mostly one, and a few function words like [*the, a, of, with*] with slightly higher frequencies. But if the syntactic structures are counted, all the nouns in a sentence contribute to the noun category, all the determiners to the determiner category, and so forth. Likewise, by subcategorizing the noun phrases into marked and unmarked types, the frequency counts are divided into two separate measures for the marked and unmarked frequency of each syntactic category. The marked and unmarked subcategorization is a way to compare different authors' patterns of use for what is salient on the one hand (as marked patterns are salient by definition) but hard to imitate on the other (as syntactic structures are fragile in memory).

B. Ground-Truth Data

The Chaski Writing Sample Database includes ten topics, listed in Table 1. The database makes cross-genre/register comparison possible for known authors who are not professional writers and produce unedited texts. With funding from the U.S. Department of Justice's National Institute of Justice, data was collected from students at a community college and a four-year college with a student body of both traditional students and returning adult students; the population provided a wide age range, males and

[48] *See id.*

[49] Chaski, *Who Wrote It?*, *supra* note 1, at 17.

females, and several races; Table 2 shows the demographics of an experiment that contrasted gender and controlled for race because race is highly correlated with some American English dialects.

Task ID	Topic
1.	Describe a traumatic or terrifying event in your life and how you overcame it.
2.	Describe someone or some people who have influenced you.
3.	What are your career goals and why?
4.	What makes you really angry?
5.	A letter of apology to your best friend
6.	A letter to your sweetheart expressing your feelings
7.	A letter to your insurance company
8.	A letter of complaint about a product or service
9.	A threatening letter to someone you know who has hurt you
10.	A threatening letter to a public official (president, governor, senator, councilman or celebrity)

Table 1: Topics in the Chaski Writing Sample Database

C. Examples of Experimental Validation Testing

With forensically feasible ground-truth data on which to run experiments testing author identification methods, ten authors were selected from the Chaski Writing Sample Database, as shown in Table 2. Each author is represented in about 100 sentences and/or 2,000 words. This was a good starting point to consider how low we could go in terms of data requirements, far less than the literary methods use, and a number that can usually be obtained in real cases. Given ten authors, there were forty-five pairwise tests of each author paired with each other author (10 * 9 / 2 = 45). At the time these experiments were run, most author identification tests were being run on two to four authors.[50] Some of the experiments reported here were first reported in my previous works.[51]

[50] *Cf.* O. de Vel et al., *Mining E-mail Content For Author Identification Forensics*, 30 ACM SIGMOD RECORD 55, 55–64 (2001); Efsthathios Stamatatos et al., *Automatic Text Categorization in Terms of Genre and Author*, 26 COMPUTATIONAL LINGUISTICS 471, 471–95 (2000); Efsthathios Stamatatos et al., *Computer-Based Authorship Attribution Without Lexical Measures*, 35

Race, Sex	**Topics by Task ID**	Author ID Number	Number of Texts	**Number of Sentences**	**Number of Words**	Average Text Size (Min, Max)
WF	1 - 4, 7, 8	16	6	107	2,706	430 (344, 557)
WF	1 - 5	23	5	134	2,175	435 (367, 500)
WF	1 - 10	80	10	118	1,959	195 (90, 323)
WF	1 - 10	96	10	108	1,928	192 (99, 258)
WF	1 - 3, 10	98	4	103	2,176	543 (450, 608)
WF Total			**35**	**570**	**10,944**	
WM	1 - 8	90	8	106	1,690	211 (168, 331)
WM	1 - 6	91	6	108	1,798	299 (196, 331)
WM	1 - 7	97	6	114	1,487	248 (219, 341)
WM	1 - 7	99	7	105	2,079	297 (151, 433)
WM	1 - 7	168	7	108	1,958	278 (248, 320)
WM Total			**34**	**541**	**9,012**	
Grand Total			**69**	**1,111**	**19,956**	

Table 2: Authors and Texts

By the time these experiments were run, empirical work with a professional statistician had shown that linear discriminant

COMPUTERS & HUMAN. 193, 193–214 (2001) [hereinafter Stamatatos et al., *Computer-Based Authorship Attribution*].

[51] Chaski, *Syntactic Analysis Method Identification*, *supra* note 31, at 3–4; Chaski, *Who's at the Keyboard?*, *supra* note 16, at 3–11.

function analysis ("LDFA") was the best statistical procedure to use for classifying an unknown document based on quantitative comparisons of two sets of known documents. LDFA is used to generate a linear function which maximizes the difference between groups; the coefficients of this function can then be used to predict the group membership of new or holdout cases.[52] In these experiments, SPSS version 13 ("Statistical Package for the Social Sciences") was used to run LDFA.

SPSS allows the user to select several variations on LDFA. The variables can be entered all together or stepwise. If the stepwise option is chosen, the user can select the number for entry or removal or use either of the defaults. The options include Wilks' lambda, F ratio, and the Mahalanobis distance. The user can also request cross-validation using a leave-one-out process. Cross-validation shows how reliable the linear function determined by the original group members is when each member is left out of the group. SPSS also allows the user to select whether prior probabilities are computed from the group sizes or not. The specific options which were chosen for each variable set are described in the experiments, as these options provide, along with different linguistic features, a series of possible experiments to run.

Experiment 1: Syntactically Classified Edge Punctuation Alone

In this experiment, only the three variables relating to syntactically classified punctuation were used. The LDFA was run with all variables entering together, prior probabilities not computed from group size, and cross-validated using leave-one-out and Wilks' lambda. Table 3 shows the cross-validation scores for each author-pair. The final row shows the average for each author. The grand average over all ten authors is 79.8% accuracy.

Author	**16**	**23**	**80**	**90**	**91**	**96**	**97**	**98**	**99**	**168**
16	X	100	94	93	92	94	92	80	93	93
23	100	X	93	93	91	93	91	67	83	92
80	94	93	X	72	75	65	81	86	71	65
90	93	93	72	X	64	66	86	75	80	47
91	92	91	75	64	X	50	58	90	54	62

[52] SPSS, SPSS 13.0 BASE USER'S GUIDE (2004).

96	94	93	65	66	50	X	75	86	70	77
97	92	91	81	86	58	75	X	80	85	85
98	80	67	86	75	90	86	80	X	82	91
99	93	83	71	80	54	70	85	82	X	86
168	93	92	65	47	62	77	85	91	86	X
Author Average	**92**	**89**	**78**	**75**	**71**	**75**	**81**	**81**	**78**	**78**

Table 3: Cross-Validation Accuracy Scores for Three Edge-Punctuation Variables

Experiment 2: Modifying the LDFA

By running the LDFA in forward stepwise mode, using Mahalanobis distance and setting F to enter at 1.84 and F to remove at 0.71 (SPSS defaults), the accuracy scores improve, over all ten authors, to 85.9%, as shown in Table 4. In Pair 91/96, none of the variables met the F levels for entering and so no analysis was run (noted as “nqv” in the table, for “no qualifying variables”). In the average for this author-pair, the sums are divided by 8 for the eight comparisons that were possible (rather than 9).

Author	**16**	**23**	**80**	**90**	**91**	**96**	**97**	**98**	**99**	**168**
16	X	100	94	100	100	100	100	70	93	100
23	100	X	87	92	91	93	91	78	83	92
80	94	87	X	83	86	70	81	86	77	71
90	100	92	83	X	64	78	93	100	80	53
91	100	91	86	64	X	nvq	83	90	69	69
96	100	93	70	78	nvq	X	75	100	82	71
97	100	91	81	93	83	75	X	100	85	92
98	70	78	86	100	90	100	100	X	91	91
99	93	83	77	80	69	82	85	91	X	86
168	100	92	71	53	69	71	92	91	86	X
Author Average	**95**	**90**	**82**	**83**	**82**	**84**	**89**	**90**	**83**	**81**

Table 4: Cross-Validation Accuracy Scores for Three Edge-Punctuation Variables (Stepwise)

Even though these three edge-punctuation variables result in an accuracy score not far below the contemporaneous results from Stamatatos et al.,[53] Baayen et al.,[54] and Tambouratzis et al.,[55] Tables 3 and 4 also show that edge punctuation may be a very good discriminator for some authors, such as 16 and 23, but a rather poor discriminator for other authors, such as 91. Further, particular author pairs are very discriminable (such as 16/23, 91/98, 168/98) while other author pairs are hardly distinguishable (such as 90/168 and 91/96), and the function is classifying near or below chance level.

Experiment 3: Adding Markedness to Syntactically Classified Edge Punctuation

In this experiment, the syntactically classified punctuation variables were combined with the marked and unmarked phrases.

[53] *See* Stamatatos et al., *Computer-Based Authorship Attribution*, *supra* note 50, at 207.

[54] *See* Harald Baayen et al., *An Experiment in Authorship Attribution*, JOURNÉES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNÉES TEXTUELLES, 2002, at 4.

[55] *See* George Tambouratzis et al., *Discriminating the Registers and Styles in the Modern Greek Language—Part 2: Extending the Feature Vector to Optimize Author Discrimination*, 19 LITERARY & LINGUISTIC COMPUTING 221 (2004).

Given earlier results, the LDFA was run stepwise, using Mahalanobis distance, and the SPSS default settings for F to enter (at 3.84) and F to remove (at 2.71) were used. The cross-validation accuracy scores are shown in Table 5.

Author	**16**	**23**	**80**	**90**	**91**	**96**	**97**	**98**	**99**	**168**
16	X	100	100	100	100	100	100	70	100	100
23	100	X	100	100	100	100	100	89	83	92
80	100	100	X	83	nvq	70	81	100	77	82
90	100	100	83	X	71	78	100	100	87	87
91	100	100	nvq	71	X	81	92	100	nvq	nvq
96	100	100	70	78	81	X	75	100	85	100
97	100	100	81	100	92	75	X	100	85	100
98	70	89	100	100	100	100	100	X	91	100
99	100	83	77	87	nvq	82	85	91	X	93
168	100	92	82	87	nvq	94	100	100	93	X
Author Average	**97**	**96**	**85**	**88**	**89**	**85**	**92**	**98**	**85**	**93**

Table 5: Cross-Validation Accuracy Scores for Markedness & Punctuation Variables

Table 5 shows that the overall accuracy rate at 90.6% with the range from 85% to 98%. Note also that for three author pairs, these variables at these default settings for the stepwise procedure did not qualify for the analysis so that no analysis was done (noted as “nqv” in the table).

Experiment 4: Syntactically Classified Edge Punctuation, Markedness, and Word Length

In this experiment, the variable set included syntactically classified punctuation, phrase markedness and average word length. The LDFA was run stepwise, using Mahalanobis distance and the default settings for F to enter and F to remove. Only one author pair had no variables qualify for the analysis under these settings.

Author	**16**	**23**	**80**	**90**	**91**	**96**	**97**	**98**	**99**	**168**
16	X	100	100	100	100	100	100	80	100	100
23	100	X	100	100	100	100	100	89	92	100
80	100	100	X	94	100	70	100	100	82	100

90	100	100	94	X	71	94	100	100	87	80
91	100	100	100	71	X	100	92	100	nvq	100
96	100	100	70	94	100	X	88	100	88	100
97	100	100	100	100	92	88	X	100	100	100
98	80	89	100	100	100	100	100	X	91	100
99	100	92	82	87	nvq	88	100	91	X	93
168	100	100	100	80	100	100	100	100	93	X
Author Average	**97**	**98**	**94**	**92**	**95**	**93**	**98**	**96**	**92**	**97**

Table 6: Cross-Validation Accuracy Scores for Markedness, Edge Punctuation, and Average Word Length Variables

Table 6 shows that the addition of word length in the variable set improves the overall accuracy rate to 95%, with individual authors' accuracy rates ranging from 92% to 98%. Note also that only one author pair was not analyzed due to "no qualifying variables" (or "nqv").

The kind of serial experimentation presented here empirically establishes a protocol, independent of any litigation, with data requirements, and known error rates that can be used in casework. One such protocol is presented below.

D. Syntactic Method Protocol using SynAID

0. Receive Q document and K documents of at least two suspects (the known authors), with approximately 100 sentences and/or approximately 2,000 words for each suspect.

1. Input Q and K documents in txt, rtf, Word format into ALIAS Documents Database.

2. Run the SynAID modules on all documents: Sentence Splitter, Tokenizer, Part-of-Speech Tagger.

3. Manually check each sentence and tag for accuracy.

4. Run the SynAID module: Markedness Subcategorizer.

5. Run the SynAID module: Punctuation Syntactic Edges Categorizer.

6. Manually check punctuation syntactic edges for accuracy.

7. Run SynAID's calculation of syntactic and punctuation quantification and average word length, for each text, normalizing so that texts of different sizes can be compared, and output the ALIAS Quantification vector for each text.

8. Input ALIAS Quantification output into SPSS (or DTReg or

Weka or R)

9. If there are a large number (50+) of K documents or multiple Q documents, run K–means clustering for internal consistency testing. If K–means clustering of K documents shows maximal subsetting, split K if needed. If K–means clustering of Q documents shows minimum subsetting, group Q.

10. Run Linear Discriminant Function Analysis on pairwise K authors, with Q held out, using leave-one-out cross-validation and equal prior probability (not set to number of documents); use SPSS default options.

11. Check classification table.

If the DFA returns high accuracy for differentiating K1 and K2,

then report classification of Q and determinative features.

If the DFA returns low accuracy for differentiating K1 and K2,

then stop. Do not use low accuracy model for classifying Q.

High accuracy is no lower than around 80% and is usually in the 90s.

Average accuracy declines for multiple authors (3 or 4) than for author pairs.

12. Check documents for nonnative English or dialectal patterns and report.

E. Admissibility

Methods in the forensic computational approach to author identification have been admitted as testimony in three trials, discussed below, with unpublished rulings of the admissibility hearings. In each of these trials, testimony based on the method was admitted without any restrictions: the expert was allowed to state a conclusion about authorship. Since these three cases did not involve any opposing experts, a fourth case involving an opposition expert that settled before trial is also discussed.

In 1998, *Erdman v. Osborne* and *Zarolia v. Osborne/Buffalo Environmental Corp.* were heard in the Circuit Court for Anne Arundel County Maryland.[56] A *Frye* hearing (a.k.a. *Frye-Reed* in

[56] Zarolia v. Buffalo Envtl., No. 1854 (Md. Ct. Spec. App. 1998); Erdman v. Osborne, No. 02C95025473 (Md. Cir. Ct. 1998), *appeal denied*, 729 A.2d 405 (Md. 1999).

Maryland[57]) was conducted, and I was examined by the attorneys and judge outside the presence of the jury. Testimony included the investigative and experimental nature of the syntactic method ("SynAID") in 1998, that the method was still being tested on a ground-truth database, and that there were current limitations still being experimentally tested. The method itself was described in detail and shown to follow standard analytical methods in linguistics and computational linguistics, as well as a common statistical procedure that was a standard technique in author identification at the time.

The court ruled that both my syntactic method for authorship identification and my analysis of second language interference were admissible without restrictions. In this case, the anonymous document could only have been written by a person in a small pool of suspects, five engineers. Writing samples from each one were analyzed using the syntactic method, and statistically, only one possible author was not differentiated from the questioned document. Also, the questioned document contained a typical first-language interference in English as a second language, i.e., the nonnative use of determiners such as [a, the]. Since many languages do not have the determiner grammatical category, using determiners such as [a, the] in the appropriate semantic places is difficult to do for nonnative speakers of English. It turned out that the lone engineer not statistically differentiated from the questioned document was a nonnative speaker of English, a native speaker of Gujarati, a language that does not have determiners.

In 2001, the United States District Court of the District of Columbia heard *Greene v. Dalton*.[58] Judge Henry Kennedy presided over a *Daubert* hearing. Testimony included reportage of validation testing results on ground-truth data, including the error rate, data requirements, and empirical standards for conducting a syntactic markedness analysis for authorship identification. Again, the SynAID method was described in detail and related to standard techniques in linguistics. I was permitted to testify about the

[57] *See* Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923) (establishing the general acceptance test used to determine the admissibility of scientific evidence); Reed v. State, 391 A.2d 364, 391 (Md. 1978) (adopting the test for admissibility established in *Frye*); *see also* MD. R. 5-702.

[58] *See* Greene v. Dalton, No. CIV.A.96-2161 TPJ, 1997 WL 33475236 (D.D.C. Oct. 3, 1997), *aff'd in part, rev'd in part*, 164 F.3d 671 (D.C. Cir. 1999).

authorship of a diary, without any restrictions on my ability to state a conclusion. The court admitted my syntactic method using SynAID without restrictions. Though the case was appealed, the diary evidence was not at issue.[59]

In 2008, the Fulton County Superior Court in Atlanta, Georgia heard *Arsenault-Gibson v. Dixon*.[60] Georgia follows the *Daubert* standard.[61] Opposing counsel filed a motion in limine regarding my syntactic method of authorship identification, so a *Daubert* hearing was conducted outside the presence of the jury. Testimony included a description of the method, error rate based on validation testing on ground-truth data outside of any litigation, and data requirements. The court rejected the motion in limine and ruled that testimony using SynAID about the authorship was admissible without restrictions.

Also in 2008, *Best Western International v. Doe*[62] was scheduled for hearing in the U.S. District Court for the District of Arizona. DLA Piper, representing Best Western International, filed a motion in limine regarding my syntactic method of authorship identification. Before a *Daubert* hearing was conducted, the experts, myself for the defendants, and Robert Leonard for the plaintiff, were deposed. The main issue was the authorship of posts on a discussion board of Best Western International franchisees, including John Doe; there were over 100 questioned posts.

My deposition testimony included a detailed discussion of the method itself, how it relates to standard methods in linguistics and computational linguistics, and the error rate and data requirements from litigation-independent testing, including the use of computational linguistics outside of litigation in Internet search engines and text classification. Regarding the particular case analysis, deposition testimony included internal consistency testing results from the known authors and document classification based on known author statistical models, including one known author with two substyles from internal consistency testing. My conclusions included both litigation-independent error rate (five

[59] *See* Greene v. Dalton, 164 F.3d 671, 674 (D.C. Cir. 1999).

[60] Arsenault-Gibson v. Dixon, No. 2004CV87715 (Ga. Super. Ct. 2008).

[61] A *Daubert* hearing is an evaluation by a trial judge on the admissibility of scientific evidence using the factors set forth in *Daubert v. Merrell Dow Pharm. Inc.*, 509 U.S. 579 (1993). *See* GA. CODE ANN. § 24-9-67.1 (2013).

[62] *See* Best Western Int'l, Inc. v. Doe, No. CV-06-1537-PHX-DGC, 2008 WL 4630313 (D. Ariz. Oct. 20, 2008).

percent) and the particular error rates associated with each statistical model for a total case-document classification error rate, as well as evidence of native language interference from one known author whose native language, Polish, has a kind of prepositional ambiguity which causes a particular linguistic interference in English. Finally, the deposition testimony included a review of academic credentials, publications, conference presentations, and previous testimony and sworn reports.

In contrast to my deposition, Leonard's deposition testimony began with the fact that neither he nor his colleagues Roger Shuy and Benji Wald had conducted any analysis of the data; instead, he testified that my method had never been heard of and could not be understood by the three linguists Leonard, Shuy, and Wald, regardless of my publications.[63] In his deposition, Leonard described my method as only a "program" that "processes text" in a way that is not transparent because he was not able to find in the text features such as marked prepositional phrases or unmarked adjective phrases. (In fact, abstract syntactic structures are not found in the text itself but in the syntactic analysis of the text.) Leonard argued that in my method I do not analyze text as a linguist but just run a program. Curiously, when Leonard described his own method, which he called sociolinguistics, he testified that he also uses computer software written by someone else to create a concordance or word list. Further, to set his own method apart from other linguists, Leonard testified that his sociolinguistics method was not forensic stylistics, even though he concurrently mentioned that he used twelve of thirteen categories listed as potential stylemarkers in the primary texts on forensic stylistics.[64] When asked about the use of his sociolinguistic method outside of any litigation, Leonard testified that it could be used as the basis for scripts for movies and television shows.

After the depositions, and due to severe restrictions by the judge on what could be presented, DLA Piper withdrew its motion in limine to exclude my testimony and SynAID method. The court

[63] *See, e.g.*, Chaski, *Empirical Evaluations*, *supra* note 16; Chaski, *Empirically Testing*, *supra* note 16; Chaski, *Syntactic Analysis Method Identification*, *supra* note 31; Chaski, *Who Wrote It?*, *supra* note 1, at 15; Chaski, *Who's at the Keyboard?*, *supra* note 16, at 1.

[64] MCMENAMIN, *supra* note 2; GERALD R. MCMENAMIN, FORENSIC LINGUISTICS: ADVANCES IN FORENSIC STYLISTICS (2002) [hereinafter MCMENAMIN, ADVANCES].

issued a summary judgment, which agreed with ninety-five percent of my report; the disagreement regarded documents I had not tested.[65] I was scheduled on a may-call list to testify, but the case settled with John Doe receiving $2 million and no gag order, an important feature to John Doe and the reason why this settlement can be reported here.

IV. Forensic Stylistics Approach to Author Identification

Forensic Stylistics is a method derived from handwriting identification, as mentioned by McMenamin[66] who quotes the standard texts of traditional handwriting identification.[67] Among the methods tested and reported in prior work was forensic stylistics as described in McMenamin.[68] McMenamin's is the only text that describes the method and the categories of "stylemarkers," which are claimed to identify each person's unique writing style. As actually practiced in the reports by Professors McMenamin, Shuy, Leonard, Coulthard, Grant, and a few other nonlinguists I have reviewed, the method consists of two steps:

1. Select stylemarkers by reading the questioned ("Q") and known ("K") documents;
2. Decide the authorship of the questioned document(s) based on the stylemarkers by listing similarities and/or differences and deciding which similarities and which differences are important or not.

The method offers:

- i. no protocol for the order of reading Q or K first, or back and forth between Q and K,
- ii. no protocol for internal consistency testing of K or Q documents, so that any number of Q documents can be

[65] *Best Western*, 2006 WL 2091695.

[66] McMenamin, *supra* note 2.

[67] *Id.* at 113–20 (reviewing the use of linguistic features by handwriting examiners in Albert S. Osborn (1910)); *see, e.g.*, James V. P. Conway, Evidential Documents (1959); Wilson R. Harrison, Suspect Documents: Their Scientific Examination (1958); Ordway Hilton, Scientific Examination of Questioned Documents (rev ed. 1982); Albert S. Osborn, The Problem of Proof (1926); Albert S. Osborn, Questioned Documents (2d ed. 1929); *see also* McMenamin, Advances, *supra* note 64, at 81–82 (attempting to distinguish the two fields of questioned document examination and forensic stylistics).

[68] *See* McMenamin, *supra* note 2.

put together, in violation of a standard forensic science principle of noncontamination;

iii. no protocol for determining the importance or "significance" of stylemarkers,

iv. no use of statistical analysis (in actual case reports); and

v. no standard reference set of stylemarkers to be reviewed in each case.

Number (v) is especially important because it means that the method allows the examiner to pick and choose stylemarkers without any predictability. This fundamental methodological flaw enables a host of problems, all rooted in subjectivity. On the one hand, it is essentially impossible to replicate a forensic stylistics analysis, while on the other hand, it is always possible to find an alternative analysis and opposing conclusion. This is the dilemma of any "pick and choose" method.

A. Litigation Dependence

Finegan documented a case in which five linguists were hired to conduct an authorship identification.[69] The five linguists each offered an opinion; each opinion used forensic stylistics to support the side which hired them. This is possible because each linguist picked stylemarkers, and each stylemarker could be deemed important or not by the linguist without any standard reference set. Finegan's report of this case demonstrates that forensic stylistics suffers from a classic case of confirmation bias being built in to a method without litigation-independent validation testing.[70]

Without litigation-independent testing, the expert battles inside litigation are inevitable. Finegan predicted that this battle of the experts would occur and that it may be a good thing:

> The expectation of expert rebuttal witnesses should contribute significantly to improvements in the quality of linguistic opinion available within the judicial system—and to justice.[71]

I would suggest that a better practice is litigation-independent validation testing, a controversial stance within the forensic stylistics community. In a recent recorded interview prior to

[69] *See* Edward Finegan, *Variation in Linguists' Analyses of Author Identification*, 65 AM. SPEECH 334 (1990).

[70] *Id.* at 339.

[71] *Id.* at 338.

deposition, Professor Leonard stated that he had “misgivings” about testing the method.[72]

B. Validation Testing

Until my research was funded by NIJ, with subsequent publications,[73] there were no known error rates for the forensic stylistics method, because none of its proponents had ever tested the method on ground-truth data, independent of any litigation, and in a blind experimental method. My prior work reports testing several authorship identification techniques, including the most common stylemarkers of forensic stylistics.[74] My prior work followed a standard blind procedure.[75] A research intern selected four female authors, around the age of forty, from the Chaski Writing Sample Database; these writing samples were typed so that no handwriting could be used to sway the analysis of the linguistic features. The intern selected one of these writing samples as the questioned document and labeled the rest of the writing samples by the numerical identifier of the writers in the database. So, the research question was, which of the four authors authored the questioned document? Each author identification technique was applied to the known writing samples first, and then the questioned document and a statistical test (χ^2 or t-test) was applied to the analytical results. The actual author of the questioned document was not revealed until all the author identification techniques were tested, and the accuracy rate for each author identification technique was then calculated.

The testing procedure in my prior work added two pieces to standard forensic stylistics: first, the method was controlled by always testing the K before the Q document, and not going back and forth between K and Q; second, a simple statistical test was applied to results.[76] So even with this strengthening of the method (from the viewpoint of scientific procedure), most of the feature categories typically selected in forensic stylistic analyses were not reliable. The actual author of the questioned document was

[72] Interview with Dr. Robert Leonard (Feb. 22, 2011).

[73] *See* Chaski, *Empirical Evaluations*, *supra* note 16; Chaski, *Who Wrote It?*, *supra* note 1.

[74] Chaski, *Empirical Evaluations*, *supra* note 16, at 3.

[75] *Id.* at 44.

[76] *See id.* at 8.

repeatedly not selected by a blind testing of stylemarker comparison.

One argument made against my prior work is that the stylemarkers were tested independently and not combined, but it is supposedly the combination of an unknown number of stylemarkers that supports the contention that each person has a unique authorial style.[77] However, anyone reading the test results could combine them, and when combined, the accuracy rate at identifying a questioned document to the real author in a pool of four authors for a combination of forensic stylistics stylemarkers is about fifty-two percent.

Forensic stylistics has very poor accuracy on ground-truth data where no one is preselected as author prior to K/Q feature selection. It is not a reliable method for authorship identification. The poor reliability of forensic stylistics, as reported in my prior article,[78] was later confirmed by validation testing using different ground-truth data by St. Vincent and Hamilton,[79] Koppel and Schler,[80] and Chaski.[81]

C. No Relationship to Standard Linguistic Methodology

Crystal[82] provided a surprisingly caustic but accurate review of McMenamin.[83]

> M[cMenamin] talks in a semistatistical way ("It is extremely unlikely that this close lexical match in profanity could be due to chance coincidence . . . .") but he does not present the statistical analysis which would make such comparisons convincing. Indeed, at several points, one wonders whether it would in principle be possible to do so, given the sample sizes, and the lack of lexical frequency norms. . . .

[77] *See id.*

[78] *See id.* at 3.

[79] *See* S. St. Vincent & T. Hamilton, *Author Identification with Simple Statistical Methods*, SWARTHMORE COLL., DEP'T OF COMPUTER SCI. (2001) (on file with author).

[80] Koppel & Schler, *supra* note 3.

[81] Chaski, *Empirically Testing*, *supra* note 16.

[82] David Crystal, Book Review, 71 LANGUAGE 381, 381–85 (1995) (reviewing MCMENAMIN, *supra* note 2).

[83] MCMENAMIN, *supra* note 2.

> The conclusion, ‘The above findings demonstrate an extraordinary level of stylistic similarity between the questioned diary and the known writings’ might in the hands of a good lawyer convince a jury, but it would not be difficult for another good lawyer to question the supposedly ‘scientific’ basis of the argument. For instance, your honor, what norms are used as the baseline for the judgments? When M says, concerning the use of the percent sign and ampersand, that ‘what . . . they have in common is their occasional use. Their use if not frequent or abnormal’, or ‘parenthesis . . . are used very frequently’, or “The semicolon . . . occurs very frequently,’ how are we to interpret these remarks? Is this linguistic SCIENCE? . . . .
>
> The problem is, after reading this book, lawyers might be forgiven for thinking that this is an orthodox account of a domain of applied stylistics. It is not. It is an account which has been tailored to meet the traditions and expectations of the legal profession . . . . It may well do a service to jurisprudence; but I am not sure that it does a service to applied linguistics.[84]

I previously described problems with the forensic stylistics method and how misleading it might be to a jury who has no concept of linguistics.[85]

Goutsos also expressed disagreement with McMenamin’s subjective assessment method.[86] In his review of McMenamin’s work for *Forensic Linguistics: The International Journal of Speech, Language and Law*, the journal of the International Association of Forensic Linguists, he shows how McMenamin’s methodology does not follow normal linguistics methodology. McMenamin evaluated as “odd” such spellings as [abit, a lot, anytime]. But when Goutsos used a typical linguistic methodology of checking for frequency in a corpus, in this case the ten-million word corpus of American English, the Bank of English Database, he found such spellings sufficient to comment: “this would imply

[84] Crystal, *supra* note 82, at 383–84.

[85] Chaski, *Who Wrote It?*, *supra* note 1; Carole E. Chaski, *Junk Science, Pre-Science and Developing Science*, NAT’L CONF. ON SCI. & L. PROC., 1999, at 97.

[86] Dionysis Goutsos, *Review Article: Forensic Stylistics*, 2 FORENSIC LINGUISTICS 99 (1995).

that careful research must precede any prescriptive judgment."[87] Indeed.

Certainly, in Professor McMenamin's defense, his later book includes a chapter in which he does consider statistics that could be used in a forensic linguistics analysis.[88] Further, he does write about a corpus he is developing.[89] But there is still a real gap between the theory put forth in the book and the method and conclusions put forth in Professor McMenamin's actual analyses and reports, as shown by Nunberg's peer review.[90]

Nunberg prepared an affidavit in which he stated:

> I believe I have a responsibility as a linguist to point out the deficiencies of Dr. McMenamin's work, which misrepresents the methods of the discipline of linguistics. . . .
>
> 1. Professor McMenamin's methods are not based on well-established theoretical principles nor are they consistent with rigorous practice in the statistical analysis of written texts. McMenamin has performed no statistical research that would give any scientific grounding to his conclusions. I would not classify McMenamin's work as bad science; rather, it is not science at all.
>
> 2. Professor McMenamin's choice of the features used in document comparison is arbitrary and subjective, and unmotivated by any empirical research; another set of features could well have been chosen that would have given very different results. His method could not pass the test of independent replicability.
>
> 3. Professor McMenamin's work is not accepted as sound science within the linguistic community. . . .
>
> The process of authorship identification is predicated on the assumption that writers may betray their individuality by certain features, and that if two documents share certain features in common there may be grounds for assuming that they have the same author. Note however that a similarity in features is not by itself a ground for assuming

[87] *Id.* at 105–06.

[88] MCMENAMIN, ADVANCES, *supra* note 64.

[89] *Id.*

[90] Statement of Geoffrey Nunberg, *In re* Marriage of Hargett, No. SDR-0017114 (Cal. Super. Ct. 2005) (on file with author).

> that two documents have the same author. That depends, rather, on how widespread these features are in the population as a whole. . . .
>
> It follows that if we have no information about the statistical frequency of various features of written texts, we can make no scientific assumptions as to whether they provide good evidence of authorship or not. . . .
>
> McMenamin has not troubled to do the work of statistical analysis necessary to teach scientific conclusions about the authorship of documents—neither in his report or in his published writings on the subject. . . .
>
> In the absence of a prior statistical analysis, McMenamin has no scientific basis for distinguishing those features of a document that are likely to be likely cues of authorship, nor does he have any grounds for assuming that the appearance of the same feature . . . in two texts offers significant evidence of common authorship. In effect, he has no way of distinguishing left-handed redheads from right-handed brunettes. Scientifically speaking, McMenamin's analyses are worthless.[91]

These reviews of forensic stylistics from other academically degreed linguists suggest two important points for judges to consider. First, forensic stylistics is not considered standard linguistics by well-established, highly regarded linguists. Second, there is certainly no general acceptance of the method, as represented by McMenamin's work, the best exposition of the method, or Leonard's testimony in the BWI deposition.

D. Admissibility

In *United States v. Van Wyk*,[92] Judge Bassler reasoned that intuition-based forensic linguistics had never been tested for its reliability, so no one knows how well or how poorly it actually works, and no one knows how much writing is required for it to work, or whether it works well or poorly at identifying authors. This lack of scientific rigor falls short of Federal Rule of Evidence 702.[93] As the court put it:

[91] *Id.*

[92] United States v. Van Wyk, 83 F. Supp. 2d 515 (D.N.J. 2000).

[93] FED. R. EVID. 702.

> Although Fitzgerald employed a particular methodology that may be subject to testing, neither Fitzgerald nor the Government has been able to identify a known rate of error, establish what amount of samples is necessary for an expert to be able to reach a conclusion as to probability of authorship, or pinpoint any meaningful peer review. Additionally, as Defense argues, there is no universally recognized standard for certifying an individual as an expert in forensic stylistics. Various judicial decisions regarding handwriting analysis, while not identical to text analysis, are instructive because handwriting analysis seems to suffer similar weakness in scientific reliability, namely the following: no known error rate, no professional or academic degrees in the field, no meaningful peer review, and no agreement as to how many exemplars are required to establish the probability of authorship.[94]

However, Judge Bassler believed that Fitzgerald's expertise in text analysis enabled him to know more about the frequency of items than the juror or judge might know.

> Unlike his opinion on authorship, Fitzgerald's expertise in text analysis can be helpful to the jury by facilitating the comparison of the documents, making distinctions, and sharing his *experience as to how common or unique a particular "marker" or pattern is.* Therefore, the Court is satisfied that Fitzgerald's testimony as to the specific similarities and idiosyncracies between the known writings and questioned writings, as well as testimony regarding, for example, *how frequently or infrequently in his experience, he has seen a particular idiosyncrasy, will aid the jury* in determining the authorship of the unknown writings.[95]

Unfortunately, Judge Bassler assumed that a person's experience as to the frequency of a previously undefined "marker" is trustworthy.[96] He assumed that a person's experience is sufficient so that he can evaluate a "marker" as idiosyncractic or

[94] *Van Wyk*, 83 F. Supp. 2d at 522 (citing United States v. Hines, 55 F. Supp. 2d 62, 69 (D. Mass. 1999); *see also* United States v. Santillan, No. CR-96-40169 DLJ, 1999 WL 1201765, at *2 (N.D. Cal. Dec. 3, 1999); Pre-Trial Transcript, United States v. McVeigh, No. 96-CR-68, 1997 WL 47724 (D. Colo. Feb. 5, 1997).

[95] *Van Wyk*, 83 F. Supp. 2d at 524 (citations omitted).

[96] *See id.*

unique. Nothing more than the expert's personal experience is offered or expected.

Judge Bassler had access to Fitzgerald's report and the book Fitzgerald relied upon, McMenamin.[97] Defense did not produce other documentation or an opposing expert, so Judge Bassler was not provided any reviews of forensic stylistics by linguists. He might have reconsidered some of his ruling if he had seen peer reviews that speak directly to the particular issue of frequency estimation in intuition-driven forensic linguistics, especially Crystal.[98]

Closely following the *Van Wyk* ruling, testimony based on forensic stylistics has been partially admitted, with the expert not allowed to state an opinion about authorship, in New Jersey[99] and Utah.[100] More in line with the scientific community's estimate of forensic stylistics, testimony based on forensic stylistics has been excluded by trial judges in California[101] and New York.[102] Testimony based on forensic stylistics has been withdrawn after a rebuttal report, depositions, affidavit, or evidence hearings in Virginia,[103] Washington,[104] and California.[105]

In a case currently under appeal, testimony based on forensic stylistics was admitted without a *Frye* hearing because the plaintiff argued that the method was not scientific and therefore not subject to *Frye*, but still presented an expert for opinion testimony.[106]

[97] McMenamin, *supra* note 2.

[98] *See* Crystal, *supra* note 82.

[99] State v. McGuire, 16 A.3d 411, 430 (N.J. Super. Ct. App. Div. 2011).

[100] United States v. Zajac, 748 F. Supp. 2d 1340, 1353 (D. Utah 2010).

[101] People v. Flinner, No. SCE211301, 2003 WL 24306950 (Cal. Super. Ct.); Beckman Coulter v. Dovatron Flextronics, No. 01CC08395 (Cal. Super. Ct. 2003).

[102] Padiyar v. Yeshiva Univ., No. 110578/05 (N.Y. Sup. Ct. filed Jan. 3, 2006).

[103] Lesnick v. Mathews, No. CL01009530-00 (Va. Cir. Ct. Nov. 21, 2003), *aff'd sub nom.* Lindeman v. Lesnick, 604 S.E.2d 55 (2004).

[104] State v. Preston, No. 02-1-03082-4 (Wash. Super. Ct. Nov. 17, 2004).

[105] *In re* Marriage of Isaacs, No. BD403783 (Cal. Super. Ct. L.A. Cnty. 2011); Hanus v. Hale, No. GIC867514 (Cal. Super. Ct. June 14, 2006).

[106] Respondent's Exceptions to Referee's Report and Brief on the Merits at 49, *In re* Masters, No. S130495 (Cal. Jan. 12, 2012).

IV. STYLOMETRIC COMPUTING APPROACH TO AUTHOR IDENTIFICATION

Stylometric disputes in literature trace their roots to the Shakespeare, Pauline, and Federalist Papers controversies. Stylometry is the measurement of style, which has a long history since the 1880s of quantifying features of written language that are easy to measure, such as sentence length, word frequency, or common words among texts. Traditional stylometric features are grounded in literary criticism, not linguistics. This kind of analysis is based on school grammar, rhetoric, and textual criticism, not linguistic theory.

With large literary datasets and the advent of computer science, stylometric computing offers more sophisticated, statistical procedures for use in comparing documents than traditional stylometry. Computer science offers, for instance, machine-learning methods for text classification. But like traditional stylometry, stylometric computing uses language features that are not grounded in linguistic theory but are easy for a computer to work with, such as character strings, words, word frequency, and common words among texts.

Recently, several researchers such as Koppel, Argamon, Juola, Chen, and their students have begun to use stylometric computing for forensic author identification.[107] In light of the best practices for forensic author identification and a recent admissibility ruling, stylometric computing currently needs to incorporate at least three of these best practices.

A. Ground-Truth Data

Ground-truth data are all too often overlooked or undervalued in stylometric computing. One intriguing study of the "writeprint" claimed a high degree of accuracy at identifying the authorship of emails, with over ninety-seven percent accuracy for English and over ninety-two percent accuracy for Chinese.[108] This impressive

[107] *See, e.g.*, Shlomo Argamon & Moshe Koppel, *A Systemic Functional Approach to Automated Authorship Analysis*, 21 J.L. & POL'Y 299 (2013); Patrick Juola, *Stylometry and Immigration: A Case Study*, 21 J.L. & POL'Y 287 (2013); Moshe Koppel et al., *Authorship Attribution: What's Easy and What's Hard?*, 21 J.L. & POL'Y 317 (2013).

[108] Jiexun Li et al., *From Fingerprint to Writeprint*, 49 COMM. ACM 9, 9–

result, however, is undermined by the fact that the dataset was not ground-truth data, as revealed by the researchers' comment about a substudy of three authors in their English dataset: "Clearly, Mike's distinct writeprint from the other two indicates his unique identity. The high degree of similarity between the writeprints of Joe and Roy suggests these two IDs might be the same person."[109] Joe and Roy's "writeprints" are almost identical. Yet it is also possible that Joe and Roy are distinct people, and the method cannot clearly recognize the difference between Joe's and Roy's documents. We will never know which explanation is correct because a dataset of ground-truth data was not used. If a ground-truth dataset had been used, if known authors were attached to one or more screennames before validation testing was begun, the accuracy of the method could have been legitimately tested.

Ground-truth data must be verified. Scraping data from the web is a fast way of collecting a lot of data, but the data are not at all easily verifiable. Koppel and his colleagues harvested a dataset of blog posts from approximately 19,000 bloggers, which is available for research.[110] The bloggers are identified by a numerical identifier, gender, age, industry, and zodiacal sign. As with any data collected from the web, there is an assumption that the screenname belongs to one person at the keyboard, but this assumption is not trustworthy, since most web-based author identification disputes focus on the facts that screennames are not reliable indicators of textual ownership. Further, ages and gender can be falsely reported and are typically not verified in any way on blog postings, or even in blog ownership.

B. Forensically Feasible Data

Traditional literary and recent computer-science-based stylometry have focused on literary texts, religious texts, and scholarly publications in science for electronic librarianship. All of the text types contain edited, rhetorically sophisticated, and highly stylized or formulaic language. These texts are also typically long, with tens of thousands of words.

10 (2006).

[109] *Id.* at 82.

[110] Jonathan Schler et al., *Effects of Age and Gender on Blogging*, AAAI SPRING SYMPOSIUM: COMPUTATIONAL APPROACHES TO ANALYZING WEBLOGS (2006).

In fact, using techniques that work well on tens of thousands of words is not at all a guarantee that it works on a few thousand (or hundred) words in an actual case of forensic author identification. Even computer tools for part-of-speech tagging that have been built on traditional "novels and newspaper" corpora will not perform well on messy, unedited texts found in forensic author identification.

C. Empirically Established Protocol

Stylometric computing methods that work on literary texts or large collections of electronic text (as in electronic librarianship) are still untested on forensically feasible data. Bringing these methods wholesale into the forensic author identification problem is not the same as empirically establishing a protocol using these methods on forensically feasible data. The stylometric computing methods must be tested on forensically feasible ground-truth data for us to know how well they really work.

Further, it is essential to make sure that the stylistic features that are being used in different components of the techniques and then subjected to the statistical multiplication rule are truly independent features. The independence of linguistic features can really only be determined by a linguistic theory, not by school grammar or literary criticism. The counting of words alone and the counting of the same words in *n*-grams are not independent counts. However, since stylometric features are so unsophisticated linguistically, these kinds of dependencies are both common and not taken into consideration in the statistical manipulations.

Finally, the number of texts required for a technique, the number of component statistical tests (with truly independent features in them, if the multiplication rule is applied), and the ability to reach a high level of accuracy on forensically feasible ground-truth data all must be established empirically before a forensic author identification method based in stylometric computing is both legally and scientifically acceptable. Fancy statistics and vague references to "research has shown" when the statistics are ill-applied and the references refer to nonforensic research could very well overwhelm a judge or jury with the aura of expertise, but it may also be seen as smoke and mirrors and not a reliable method when the smoke clears.

D. Admissibility

In *United States v. Fresenius* in the District Court for the Western District of Texas,[111] the court ruled in favor of Fresenius's motion in limine to exclude stylometric computing testimony regarding the authorship of medical records. The proffered method focused on words, a standard stylometric analytical level. The statistical techniques included the Bernoulli mixture method. Yet even with a standard word-based stylometry and sophisticated statistical analysis, Judge Martinez ruled the testimony inadmissible because the expert, a professor of computational linguistics at the University of Texas, whose credentials were duly noted as impressive, could not offer any error rate or any verification of his method, while also maintaining that his method was 100% accurate. Judge Martinez's ruling warns us that sophisticated statistical analysis does not replace the need for empirically established protocols with known error rates through validation testing of each method on forensically feasible ground-truth data.

VI. CONCLUSION

Some scholars cast these three approaches, in a binary distinction, as intuition versus algorithm or nonquantitative versus quantitative.[112] From this perspective, forensic stylistics (the nonquantitative, intuitive approach) stands in contrast to forensic computational linguistics and stylometric computing (both of which are algorithmic and quantitative). I would suggest that there are two other binary distinctions to be considered in evaluating current approaches to forensic author identification.

First is the role of linguistics: is the approach linguistics or not? Forensic computational linguistics is grounded in linguistic theory, implements linguistic analysis in software, and uses standard linguistic methodology not only for analytical techniques but also for data collection and research methodology. Neither forensic stylistics nor stylometric computing is grounded in linguistic

[111] United States *ex rel.* Gonzalez v. Fresenius Med. Care N. Am., 748 F. Supp. 2d 95 (W.D. Tex. 2010), *aff'd sub nom.* Gonzalez v. Fresenius Med. Care N. Am., 689 F.3d 470 (5th Cir. 2012).

[112] *See, e.g.*, Lawrence M. Solan, *Intuition Versus Algorithm: The Case of Forensic Authorship Attribution*, 21 J.L. & POL'Y 551 (2013).

theory. Instead, both forensic stylistics and stylometric computing are grounded in conceptions of language that are common in prescriptive grammar and literary criticism or focused on naïve conceptions of language as a list of words or a list of function words. So considering the "linguistics" in forensic linguistics, of which author identification is a primary task, forensic computational linguistics employs standard linguistics, while forensic stylistics and computer science neither use linguistics in analytical techniques nor theoretical underpinnings.

Second is the role of research in the approaches. In order for the *Daubert* factors to be met, litigation-independent validation testing on forensically feasible "ground-truth" data must be conducted. Forensic computational linguistics has met this challenge directly through the use of forensically feasible "ground-truth" datasets such as the Chaski Writer Sample Database. Independent of any litigation, validation tests have been conducted, as reported earlier in this paper. These tests have been run on forensically feasible data—that is, documents which are short, in several types of genre and register, and without any correction to grammar, spelling, or prescriptive conventions about writing. Further, the data are ground-truth data, where the authorship of each document is known; there is no possibility that someone else was using a screenname or posting blogs under a pseudonym. Finally, the validation test research has resulted in a known protocol for what is needed to apply the forensic computational linguistic methods; the test results empirically limit the amount of data required. It is hoped that both forensic stylistics and stylometric computing will conduct the kind of research that forensic computational linguistics performs, so that reliable methods of forensic authorship identification can be offered to our courts.