Who's At The Keyboard?
Authorship Attribution in Digital Evidence Investigations

Carole E. Chaski, Ph.D.
Institute for Linguistic Evidence, Inc.

Abstract

In some investigations of digital crime, the question of who was at the keyboard when incriminating documents were produced can be legitimately raised. Authorship attribution can then contribute to the investigation. Authorship methods which focus on linguistic characteristics currently have accuracy rates ranging from 72% to 89%, within the computational paradigm. This article presents a computational, stylometric method which has obtained 95% accuracy and has been successfully used in investigating and adjudicating several crimes involving digital evidence. The article concludes with a brief review of the current admissibility status of authorship identification techniques.

Section 1: The Need and Available Methods

In the investigation of certain crimes involving digital evidence, when a specific machine is identified as the source of documents, a legitimate question is, "Who was at the keyboard when the relevant documents were produced?" For example, consider the following scenarios, drawn from actual cases.

1. A government employee wrote e-mails to his supervisor in which he disparaged her racial heritage. After he was terminated for cause, he sued the federal government, claiming that his workspace cubicle had been open, allowing any of his co-workers to author the e-mails on his computer and send them from his computer without his knowledge.

2. A young, healthy man was found dead in his own bed by his roommate who notified the police. When the autopsy results showed that he died by injection, his death was investigated as a potential homicide. During the investigation, the roommate gave the police suicide notes which he found on the home computer. These had never been printed or discovered before the death.

3. A civilian intern with a military research laboratory kept an electronic journal of her relationship with her supervisor. As her internship came to a close, she claimed that her relationship with her supervisor was not mutually consenting and that he had raped her. When the intern's work computer was searched, the journal was discovered. The intern claimed that during the time she had not had access to the work computer or the journal, her supervisor had edited the journal to agree with his version of the events.

Apparently, the question of who was at the keyboard has arisen in many other investigations as well. Chuck Davis, National Center for Forensic Sciences, Digital Evidence Division, is a former special agent with the Air Force Office of Special Investigations and the NASA Computer Crime Division, and a former agent with the Colorado Bureau of Investigation. Davis confirms that the authorship question has been raised in a variety of computer crime investigations ranging from homicide to identity theft and many types of financial crimes (Davis, personal communication). According to Davis, a significant number of investigations require that, in order to prove a case, investigators must put the hands of a particular suspect on the keyboard. During a large number of bomb and death threat investigations, Davis was often faced with instances where multiple people had unrestricted physical access to a computer system in questions. Usually these situations involved family members, college roommates, members of the same office, and similar circumstances. In one example from Davis' experience, an elected official had received a number of both written and e-mailed death threats from within the agency e-mail system. It was eventually determined that the official herself had been authoring the threats and was never in any danger at all. However, Davis used traditional investigative techniques, such as interviews, hand-writing analysis and others, to determine the true author of the threats.

In another investigation, Davis received an allegation of racial intimidation occurring between employees in a government agency. A barrage of e-mails rife with ethnic slurs had been exchanged over several weeks' time through anonymous e-mail services. Davis again had to rely on standard police work to track down which employees were involved and bring the case to a successful conclusion.

Another type of investigation where the "whose hands are on the keyboard" is a question of great importance is the sexual exploitation of children. Davis investigated a case where a young girl had been involved in a series of sexually explicit exchanges via an instant messenger system. Upon investigation, the perpetrator was tracked to the home of a prominent local physician. However, the case took a twist when Davis determined that the doctor's 13-year-old son had been using his father's account to have conversations with the girl.

Obviously, any method for determining authorship must work in conjunction with the standard investigative and forensic techniques which are currently available, as demonstrated by the examples provided by Davis. Determining who was at the keyboard can be approached through several avenues: biometric analysis of the computer user; qualitative analysis of "idiosyncrasies" in the language in questioned and known documents; and quantitative, computational stylometric analysis of the language in questioned and known documents.

First, the biometric approach has focused on actual keyboard stroke dynamics (Gupta, Mazamdur & Rao 2004). A software driver associated with the keyboard records the user's rhythm in typing; these rhythms are then used to generate a profile of the authentic user. Although this solution is non-linguistic, linguistic characteristics such as the phonotactics of each language and language family (e.g., the presence of word-

initial [tl] in Greek compared to the absence of word-initial [tl] in English) may actually affect how such profiles are generated. Given the focus of this article, however, this line of inquiry will not be pursued further.

Second, the qualitative approach to authorship assesses errors and "idiosyncrasies" based on the examiner's experience (McMenamin 1993, 2001, Foster 2000, Ollson 2004). This approach, known as forensic stylistics, could be quantified through databasing, as suggested by McMenamin (2001), but at this time the databases which would be required have not been fully developed. Without the databases to ground the significance of stylistic features, the examiner's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias. Another approach to quantifying is counting particular errors or idiosyncrasies and inputting this into a statistical classification procedure. When the forensic stylistics approach was quantified in this way by Koppel and Schler (2000), using 100 "stylemarkers" in a Support Vector Machine (Vapnik 1995) and C4.5 (Quinlan 1993) analysis, the highest accuracy for author attribution was 72%.

A third approach, stylometry, is quantitative and computational, focusing on readily computable and countable language features, e.g. word length, phrase length, sentence length, vocabulary frequency, distribution of words of different lengths. (For an overview of this see Holmes 1996). Examples of this approach using various statistical procedures include deVel et al. (2001), Stamatatos et al. (2001), Tambouratzes et al. (2004), and Baayen et al. (2002). Using Support Vector Machine, de Vel et al. (2001) obtained accuracy rates which were very high (100%) or very low (46%), depending on the author pairs. Stamatatos et al. (2001), Tambouratzes et al. (2004), and Baayen et al. (2002) each used discriminant function analysis; their reported accuracy results range from 87% to 89%. Using neural networks (Haykin 1999), Diri and Amasyali (2003) obtained 84% accuracy. All of these studies used a small number of authors (4, 10, 5, 8, 18, respectively) and a diverse number of total texts per author (1259, 30, 1000, 72, 270 respectively). This article presents a quantitative method, within this same computational stylometric paradigm, which uses standard syntactic analysis from the dominant paradigm in theoretical linguistics over the past forty years. The syntactic analysis method (Chaski 1997, 2001, 2004) has obtained an accuracy rate of 95%. The primary difference between the syntactic analysis method and other computational stylometric methods is the syntactic method's linguistic sophistication and foundation in linguistic theory. Typical stylometric features such as word length and sentence length are easy to compute even if not very interesting in terms of linguistic theory, but the more difficult to compute features such as phrasal type are also more theoretically grounded in linguistic science and experimental psycholinguistics.

In any pattern-matching task, the basic problem is finding the right feature sets to input into the right classification procedure. The task of achieving feature-algorithm optimality within the forensic setting is hampered by the fact that data may be extremely small so that variables must be selected with special care in order to avoid "the curse of dimensionality," i.e., having so many more dimensions in the feature set than cases. In this work, I have focused specifically on reducing the number of variables so that a

method can operate on a very limited number of documents; in related work, Chaski and Chmelynski (2005a, 2005b) have focused on decomposing the documents so that the method can operate on a larger number of variables even with a limited number of documents.

Section 2: Data, Method and Result

The Data: Authors and Texts

Based on sociolinguistically-relevant demographics and the amount of text, ten authors were drawn from Chaski's Writing Sample Database, a collection of writings on particular topics designed to elicit several registers such as narrative, business letter, love letter, and personal essay (Chaski 1997, 2001). Sociolinguistically-relevant demographics include sex, race, education and age. These demographic features can be used to define dialects. Controlling for these features tests the ability to differentiate authors at an individual rather than group level. Although this dataset was not as tightly constrained as the dataset in Chaski (2001), because it includes both men and women and a wider age range, this dataset has been controlled for race and education. The five women and five men are all white adults who have completed high school and up to three years of college at open-admission colleges. The authors range in age from 18 to 48. The authors all have extensive or lifetime experience in the American English, Delmarva dialect of the mid-Atlantic region of the United States. The authors are "naïve writers" (in terms of Baayen, et al. 2002) with similar background and training. The authors volunteered to write, wrote at their leisure, and were compensated for their writings through grant funding from the National Institute of Justice, US Department of Justice.

Another control for the dataset is the topic. Controlling the topic tests the ability to differentiate authors even though they are writing about the same topic. The authors all wrote on similar topics, listed in Table 1.

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 Writing Sample Database

Further, the author selection took into consideration the quantity of writing which the authors had produced. Authors who met the sociolinguistic demographics, but produced only three documents were not included in this dataset lest the lack of data produce misleading results. In order to have enough data for the statistical procedure to work, but in order to make this experiment as forensically feasible as possible, the number of documents for each author was determined by however many were needed to hit targets of approximately 100 sentences and/or 2,000 words. One author needed only 4 documents to hit both targets, while two authors needed ten documents. Three authors needed 6 documents to hit the sentences target, but only one of these three authors exceeded the words target. The exact details are shown in Table 2: Authors and Texts.

Race, Gender	Topics by Task ID	Author ID Number	Number of Texts	Number of Sentences	Number of Words	Average in Words 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

The Feature Set[1]

Each text was processed using ALIAS, a program developed by Chaski (1997, 2001) for the purpose of databasing texts, lemmatizing[2], computing lexical frequency ranking[3], calculating lexical, sentential and text lengths, punctuation-edge counting, Part-Of-Speech-tagging[4], n-graph and n-gram sorting[5], and markedness subcategorizing. ALIAS is thus able to provide a large number of linguistic variables. In this study, however, only three types of variables are used: punctuation, syntactic and lexical.

Chaski (2001) showed that syntactically-classified punctuation had a slighter better performance than simple punctuation marks for discriminating authors while preserving intra-author classification. Authors may share the same array of marks, but the placement of the marks appears to be what matters. This approach to using punctuation as an authorial identifier –syntactically-classified punctuation – is very different from the approach – simple punctuation marks – advocated by questioned document examination (Hilton, 1993), forensic stylistics (McMenamin 2003), as well as the other computational stylometric studies discussed earlier. In simple punctuation approaches, the punctuation marks themselves, such as commas, colons, exclamation points, etc., are counted. In the syntactically-classified punctuation approach, the marks (no matter what they specifically are) are counted by the kind of boundary or edge which the punctuation is marking.

After each text is automatically split into sentences, the user interacts with ALIAS to categorize punctuation within each sentence by the syntactic edge which it marks. These syntactic edges are the clause, the phrase, and the morpheme (word-internal). The end-of-clause (EOC) marks may be commas, semi-colons, hyphens; the particular marks are not counted separately, but any and every EOC mark is counted. Again, the phrase edge may be marked by hyphens or commas; what is counted is the marked EOP edge. Word-internal edges typically are morphemic edges, a morpheme being a minimal unit of meaning. For example, the word [re-invent] includes two morphemes, [re] and [invent], and the hyphen marks the edge of the morpheme [re], just as an apostrophe typically marks the morphemic, word internal edge in [don't] and [can't]. The morphemic edges which are marked by some punctuation (hyphen or apostrophe) are counted. ALIAS then exports these syntactically-classified punctuation counts to a spreadsheet.

Chaski (1997) showed that syntactic markedness could preserve intra-author identification while performing inter-author discrimination. Markedness is the basic asymmetry in language which pervades the binary substructure of linguistic signs; even though language is structured for contrasts, the contrastive items are not equally

[1] A U.S. patent is pending for the variables and method of authorship attribution presented herein.
[2] Lemmatizing converts inflected word forms (such as plurals) into base word forms (such as dictionary look-up forms).
[3] Lexical frequency ranking refers to ordering words from the most frequently-used to least frequently-used in a text.
[4] Part-Of-Speech tagging labels each word by its grammatical function such as noun, verb, preposition and so forth.
[5] N-graph refers to a specific number (n) of letters in sequence; N-gram refers to a specific number (n) of parts-of-speech labels or words in sequence. Once these sequences are found, they can be sorted by similarity.

interchangeable. For example, the binary contrast of the concept [age] is lexicalized in English as [young] / [old]. But the binary distinction between [young] / [old] is not symmetrical as shown by the fact that these two terms are not interchangeable. When we are inquiring about age in English, we ask [how old are you?] for the unmarked use, while we can, in a marked use, ask [just how young are you?].

For syntactic structures, the unmarked contrast is the most common and often the most easily parsed, while the marked contrast is typically less frequent and sometimes more difficult to parse because it can pose several different parsing attachments. For example, the head-position of the noun, in universal terms, can be either initial or final (the binary contrast). This head-position parameter distinguishes English and Spanish, since in simple noun phrases, the English noun will be in final position, after any modifiers, while the Spanish noun will be in initial position, before any modifiers, as shown in (1).

(1) English: white house

ADJECTIVE NOUN => NOUN PHRASE [NP]

HEAD

(2) Spanish: casa blanca

NOUN ADJECTIVE => NOUN PHRASE [NP]

HEAD

The unmarked noun phrase in English is head-final while the unmarked noun phrase in Spanish is head-initial. But in both English and Spanish, the marked variants are possible. In English, noun phrases which are not head final include head-medial structures such as NP [Determiner-phrase Adjective-Phrase Noun Prepositional-Phrase] (such as 'the white house on the corner') and NP[Determiner-Phrase Adjective-Phrase Noun Complementizer-Phrase] (such as 'the white house that your brother bought last summer') as well as the more rare head-initial structure NP [Noun Adjective-Phrase] (such as 'forest primeval'). Each syntactic head has its own markedness properties. While head-position is marked or unmarked for nouns, a semantic feature of predicative or attributive is the markedness contrast for adjectives, the syntactic property of recursion is the markedness contrast for prepositions, and syntactic attachment to verbal or non-verbal heads is the markedness contrast for modifiers. For an overview of markedness theory, see Battistella (1990). The markedness contrasts related to authorship identification have been influenced both by the general theory of markedness as well as empirical tests beginning with Chaski (1997).

The following sentences in (3) have been subcategorized for syntactic markedness in (4).

(3) Markedness pervades all levels of language, from phonetics to morphology to syntax to semantics to discourse. Actually, markedness may pervade many other forms of cultural artifact. We humans seem aware of what is usual and common and entropically perk up when we encounter what is unusual or slightly odd or somewhat complex.

(4) Marked Noun Phrases: all levels of language, many other forms of cultural artifact
Unmarked Noun Phrases: markedness, language, phonetics, morphology, cultural artifact

Marked Verb Phrases: may pervade
Unmarked Verb Phrases: pervades, seem, is, encounter

Marked Adjective Phrases: aware of, usual, common, slightly odd, somewhat complex
Unmarked Adjective Phrases: cultural

Marked Prepositional Phrases: From phonetics to morphology to syntax, of what is usual
Unmarked Prepositional Phrases: of language, of cultural artifact

Marked Modifier Phrases: actually, entropically
Unmarked Modifier Phrases: slightly, somewhat

After each word is tagged for its part-of-speech, ALIAS searches and sorts syntactic head patterns and flags the pattern exemplars as either marked or unmarked. These flags can be checked by the user. ALIAS then counts the marked and unmarked exemplars for each syntactic head, collapses them into two variables (marked XP and unmarked XP), and outputs these counts to a spreadsheet for statistical analysis.

Finally, one lexical variable was included. Following the lead of Tambouratzis et al. (2004) and many other stylometric studies, average word length for each document was computed. All words, both function and content words, were included in this computation.

In sum, as listed below, there are three syntactically-classified punctuation variables, two syntactic markedness variables and one lexical variable.

The Statistical Procedure

SPSS version 13 (Statistical Package for the Social Sciences) was used to run linear discriminant function analysis (DFA). Discriminant function analysis 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.

SPSS allows the user to select several variations on DFA. 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 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. The options for cross-validation include Wilk's lambda, F ratio, and the Mahalanobis distance. SPSS also allows the user to select whether prior probabilities are computed from the group sizes or not. In this experiment, the DFA was run stepwise, with SPSS default settings for F to enter and F to remove. Leave-one-out cross-validation was selected using Mahalanobis distance. Prior probabilities were computed based on group size. Under these settings, only one author pair (91-99) had no variables qualify for the analysis.

Given 10 authors, there were 45 pairwise tests of each author paired with each other author (10*9/2 = 45). Table 3 shows the overall accuracy rate to 95%, with individual authors' accuracy rates ranging from 92% to 98%.

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	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**	**94**	**92**	**97**

Table 3: Cross-Validation Accuracy Scores for the Chaski Feature Set

Section 3: Applying the Syntactic Analysis Method to Casework

Validation studies provide the kind of information required for an evidence admissibility hearing, but actual casework often requires a slightly different angle. For instance, in casework attorneys and investigators want to know not just whether the method has a high accuracy rate/low error rate, but also the probability associated with an identification. In other words, how likely is it that the questioned documents are different from one suspect? How likely is it that the questioned documents belong to another suspect? Discriminant function analysis, like many other classification procedures, does not really provide such a probability directly. The probability associated with Wilks' lamba can indicate whether the discriminant function is or is not significant, but a significant Wilks' lambda can sometimes occur with a poor cross-validation rate. One way to provide a probability value in actual casework is to analyze the discriminant scores from different authors with a standard t-test. Given the 100% cross-validated

accuracy rate of the discriminant function analysis of authors 16 and 23, for instance, the Wilks' lambda was highly significant, p =.001. The discriminant-scores, or distance measures between the documents of 16 and 23, are listed in Table 4.

Author	Document	D-Score
16	1	1.215
16	2	2.939
16	3	.784
16	4	.065
16	5	2.795
16	6	.209
23	1	-1.228
23	2	-1.659
23	3	-1.946
23	4	-.940
23	5	-2.233

Table 4: D-Scores for Authors 16 and 23

When these scores were analyzed with an independent samples t-test for equality of means, the probability that these d-scores come from the same population is very small, p = .001. The t-test result supports the cross-validation result and the conclusion that documents from author 16 and author 23 significantly differ from each other and can be accurately differentiated.

Further, attorneys and investigators are also rightly concerned about communicating information to the judge and/or jury in a simple, clear way. The discriminant scores can be graphed to show the clear separation between authors. Figure 1 shows the d-scores of 16 and 23, clearly indicating the separation between 16 and 23.

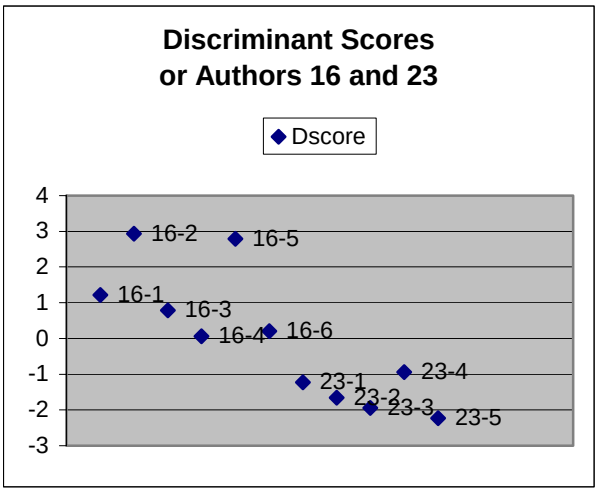

Figure 1: Graph of D-Scores for Authors 16 and 23

Chaski (2005-pending) reports a recent case involving authorship of non-digital documents in which the syntactic analysis method described earlier was applied; this report includes more details than presented here.

Admissibility

The answer to this question –who was at the keyboard? – might be used to generate investigative leads, to produce leverage in negotiating a settlement ,or to present admissible evidence in a trial. The expectation of scientific validation increases as one moves from investigative lead to leverage to admissible evidence. In the federal courts and states which have accepted the Daubert criteria, scientific validation of a forensic technique involves peer review, experimental results determining error rates, and standardized operating procedures, and it is expected that such indicia of science will be presented to the court. In states which have maintained the Frye criterion, scientific validation of a forensic technique relies on the scientific community's acceptance of a method as the primary indicator that a method is reliable, falsifiable and within the realm of standard scientific endeavor. Whether the Daubert or Frye criteria are in play from the legal perspective, from the scientific perspective, producing good "normal science" in the Kuhnian sense will automatically meet legal criteria, because good science is empirical, operationalizes hypotheses for falsifiablity, requires replicability, seeks knowledgeable criticism and review.

Complaining that the Daubert criteria for scientific and technical evidence is wrong and needs to be changed, as proponents of forensic stylistics Olsson (2004) and McMenamin (2002) do, just delays the inevitable research that needs to be done. Nor does such complaining persuade judges that forensic stylistics is engaging in normal scientific activity. Ignoring the legal responsibilities of forensic science, as the proponent of text analysis (a twin to forensic stylistics) Foster (2000) does, imperils the credibility of author attribution, making it appear as nothing more than academic posturing. (See the case Hatfill v. Foster, Conde Nast Publications, Reader's Digest Association, et al currently filed in United States District Court, Alexandria, VA.) The forensic stylistics/text analysis method has been restricted to limited admissibility by a federal judge in a Daubert hearing and it has been excluded from testimony completely in two evidence hearings in the state of California, which holds to the Kelly-Frye standard (U.S. v. van Wyk, New Jersey 2000; California v. Flinner, San Diego, CA 2003; Beckman Coulter v. Dovatron/Flextronics, Santa Monica, CA 2003). In at least two other cases in California, courts have followed the van Wyk decision and admitted forensic stylistics testimony without any expert conclusion regarding authorship.

The syntactic analysis method of authorship identification (Chaski 1997, 2001) has been scrutinized by a federal judge in a Daubert hearing and its evidence has been allowed into trial with full admissibility (Green v. Dalton/U.S. Navy, District of Columbia). In Maryland, which holds to the Frye standard, testimony based on the syntactic analysis method was admitted, including the expert's opinion as to authorship (Zarolia v. Osborne/Buffalo Environmental Corp, Annapolis).

The syntactic analysis method, with some variations, has been used in the cases numerically listed earlier.

1. The dismissed employee withdrew his suit against the government after the questioned emails were identified as his own writing based on the syntactic patterns.

2. The roommate was arrested, charged, and tried for first degree murder. Although he pled not guilty, he confessed on the witness stand to writing the suicide notes.

3. The supervisor was cleared of the rape charge in a military trial and later faced similar charges in a civil trial. The supervisor was again cleared of charges in the civil trial, part of which was testimony demonstrating through the syntactic patterns that the electronic journal was identifiable with the intern's own writing.

Section 4: Related Work and Conclusion

In related work, Chaski and Chemylinski (2005a-pending) have developed a method for decomposing the data into smaller chunks so that a larger set of variables can be used for the discriminant analysis. The overall accuracy rates are congruent with the results in this report at 95.7%. Chaski and Chemylinski (2005b-pending) have also obtained similar results using these variables with logistic regression. In future work, we plan to run experiments using additional authors and also to explore additional statistical procedures including support vector machines as suggested in deVel et al (2001) and Koppel and Schler (2001).

Finally, these experiments have demonstrated the possibility of a reliable method for determining authorship which uses linguistically defensible units of analysis and is forensically feasible in terms of the brevity and scarcity of texts. Because this particular method obtains a high degree of reliability when it is subjected to a cross-validated statistical procedure, it really is possible to determine who was at the keyboard.

About the Author

Carole E. Chaski earned her AB magna cum laude in Greek and English from Bryn Mawr College, MEd in Psychology of Reading from University of Delaware, and MA and PhD in Linguistics from Brown University. She has consulted with law enforcement, attorneys, and private individuals since 1992. After a Visiting Fellowship with the National Institute of Justice, Dr. Chaski founded the Institute for Linguistic Evidence, Inc, a non-profit corporation dedicated to research and service in the forensic application of

linguistics, of which she is currently Executive Director. She can be contacted at 302-856-9488 and cchaski@LinguisticEvidence.org.

References

Baayen, H., van Halteran, H., Neijt, A., Tweedie, F. (2002). "An Experiment in Authorship Attribution." Journees internationales d'Analyse statistique des Donnees Textuelles 6.

Battistella, E. (1990). Markedness: The Evaluative Superstructure of Language. Albany, State University of New York Press.

Chaski, C. E. (1997). "Who Wrote It? Steps Toward A Science of Authorship Identification." National Institute of Justice Journal. September:15-22.

Chaski, C. E. (2001). "Empirical Evaluations of Language-Based Author Identification Techniques." Forensic Linguistics 8(1): 1-65.

Chaski, C. E. (2004). “Recent Validation Results for the Syntactic Analysis Method for Author Identification.” International Conference on Language and Law, Cardiff, Wales.

Chaski, C. E., and Chmelynski, H. J. (2005a-pending publication). "Testing Twenty Variables for Author Attribution by Discriminant Function Analysis." Ms.

Chaski, C. E., and Chmelynski, H. J. (2005a-pending publication). "Testing Twenty Variables for Author Attribution by Logistic Regression." Ms.

deVel, O., A. Anderson, M. Corney, G. Mohay (2001). “Multi-topic E-Mail Authorship Attribution Forensics.” ACM Conference on Computer Security-Workshop on Data Mining for Security Applications. Philadelphia, PA.

Diri, B. and Amasyali, M. F. (2003). “Automatic Author Detection for Turkish Texts.” ICANN/ICONIP. Available at www.ce.yildiz.edu.tr/mygetfile.php?id=265.

Foster, D. (2000). Author Unknown: On the Trail of Anonymous. New York, Henry Holt and Co.

Gupta, G., Mazumdar, Chandan, Rao, M.S. (2004). "Digital Forensic Analysis of E-Mails: A Trusted E-Mail Protocol." International Journal of Digital Evidence 2(4): 1-11.

Haykin, S. (1999). Neural Networks. New York, Prentice Hall.

Hilton, O. (1993). Scientific Examination of Questioned Documents. Boca Raton, Florida, CRC Press.

McMenamin, G. R. (2003). Forensic Linguistics: Advances in Forensic Stylistics. Boca Raton, Florida, CRC Press.

Olsson, J. (2004). Forensic Linguistics: An Introduction to Language, Crime and the Law. New York: Continuum.

Mosteller, F., and Wallace, D. L. (1984). Applied Bayesian and Classical Inference: The Case of the Federalist Papers. New York, Springer-Verlag.

Quinlan, J. (1993). C4.5: Programs for Machine Learning. San Mateo, CA, Morgan Kaufmann Publishers.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). "Automatic Text Categorization in Terms of Genre and Author." Computational Linguistics 26(4): 471-495.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). "Computer-Based Authorship Attribution Without Lexical Measures." Computers and the Humanities 35: 193-214.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., Tambouratzis, D. (2004). "Discriminating the Registers and Styles in the Modern Greek Language -- Part 2: Extending the feature Vector to Optimize Author Discrimination." Literary & Linguistic Computing 19(2): 221-242.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. New York, Springer-Verlag.