The Computational-Linguistic Approach to Forensic Authorship Attribution

Carole E. Chaski, PhD
Institute for Linguistic Evidence, Inc.
Georgetown, DE 19947 USA

This article examines the diversity of methods in authorship attribution through a lens which focuses attention on a single common element. The current state of authorship attribution study is spread throughout so many academic and non -academic disciplines that it is nigh impossible to describe all of the various assumptions about language and authorship. The disciplines involved in authorship attributions range over Classics, Biblical exegesis, Paleography, Communication and Rhetoric, English literary criticism, Handwriting examination, General Linguistics, Sociolinguistics, Computational Linguistics, Statistics, and Machine Learning. Given the breadth of this list, it is no wonder that the current state of authorship attribution appears to be a jumbled mass of multiple contrasts.

For instance, methods can be characterized as
qualitative vs. quantitative, --are linguistic features noted or counted?
prescriptive vs. descriptive, --are the documents evaluated by an explicit or implicit standard or described by any and all contents?
analyst-based vs. machine-based, --does the method require an analys t's particular subjective expertise or can the method be objectively operationalized and automated?
theory-based vs. practice-based –does the method refer to and rely on a theory of language or the analyst's individual experience with language or texts?

These binary contrasts are not mutually exclusive. Two methods can be prescriptive and practice-based, where one is quantitative and the other qualitative or one is analyst-based and the other machine-based. Obviously, it is an exciting time to be researching authorship attribution because there are so many ways to formulate a method, even if it is a bit overwhelming to try to comprehend such diversity in methods. But just as a kaleidoscope can be positioned so that some unifying pattern finally emerges, I suggest that such a twist in perception exists for authorship attribution as well.

Authorship attribution –from any discipline—is essentially a pattern recognition problem. As analysts, we have patterns of linguistic behavior in documents of unknown authorship and documents of known authorship and we are trying to estimate how similar or dissimilar the two sets of patterns are from each other. Different methods define how we come up with the patterns of linguistic behavior –whether we extract them by a computer program or our linguistic instincts. Different methods define what patterns of linguistic behavior are considered –whether these patterns include orthography, words, function words, grammatical errors or part of speech tags. Different methods define how we estimate similarity and dissimilarity—whether we rely on statistical analysis, machine learning algorithms or subjective judgments.

When we take this global perspective on authorship attribution research , and focus on the idea that every authorshi p attribution method is a pattern recognition method, then the core issue for each method is empirical validation. Proponents of each method bear responsibility for testing it, independent of any litigation, to determine how well the method correctly attri butes authorship under a variety of realistic and controlled conditions. Seeing is believing!

The standard method for empirically validating any classification method is very logical and simple. The analyst starts with a set of objects, in our case, texts, whose authorship (or gender or topic) class is already known and undisputed. These objects are typically divided into two known classes (author A and author B, male or female gender, or on-topic and off-topic). The linguistic features (forensic stylistic 'stylemarkers', or words or sentence length or paragraphs per document or n -grams) are extracted from each text; their presence and/or frequency is tabulated. The feature data are analyzed by a statistical or machine learning algorithm which builds a model of the data classified into two classes and outputs a confusion matrix. The model shows the separability of the data into two classes. The confusion matrix lists how many texts of a known category were classified as the correct class and how many were not. So, for instance, if author A had 5 texts and author B had 5 texts, how separable are the A and B texts, and in particular, how many of the 5 A texts were classified correctly as A texts, and how many were confused with B texts? How many of the 5 B tex ts were classified correctly as B texts, and how many were confused as A texts? The confusion matrix gives us information about the "hits" (the correct classifications) and the "misses" (the incorrect classifications) so that we get an accuracy score.

When the amount of texts for the two classes is small, as in the typical forensic setting, cross-validation is used to test how accurate the classification algorithm is performing. Cross-validation means that a portion of the original data (i.e. some of the ten texts of A and B) is withheld from the model -building and then classified by the model. Leave-one-out cross-validation means that each text is withheld from the model building, and its class is predicted by that model; N-fold cross-validation means that a certain nth, such as one-tenth or one -fifth, of the data is withheld from the model building and then classified by that model. If there are ten documents, and we use leave -one-out cross-validation, then ten models are built, and each document is class ified using a model it did not help to build. The hits and misses for each model are averaged for the final cross-validated accuracy score.

Since A and B each has the same number of texts, then any text has an equal or 50% chance of being classified correctly as either A or B. If the algorithm, with a particular set of linguistic features, returns a cross-validated accuracy score of 100%, then we know that A and B can be classified correctly much higher than the base rate of 50%. But if the algorithms returns a cross-validated accuracy score of 50%, then we know that A and B can only be classified at the base rate or chance level of 50%.

The cross-validated accuracy scores tell us two things: first, if our particular linguistic features are any good at distinguishing these two authors using this particular classification algorithm, and second, if this particular classification algorithm is any good at distinguishing these two authors with this particular set of linguistic features.

The cross-validated accuracy score does not tell us how likely it is that author A wrote a particular text. This is so important that it deserves repeating. The cross -validated accuracy score does not tell us about the probability that a certain author wrote a particular text. The cross-validated accuracy score is not a probability; it is only a record of hits and misses for the classification procedure. It only tells us about the empirical reliability of the method. It is important to use the method with as much data and as many conditional variations of the data as possible so that we can see how reliable the method is overall, over a large set of authors, over a small set of documents, over a wide range of topics, and so forth, befitting the forensic setting.

Given this perspective, this article addresses three broad approaches to forensic authorship attribution: (1) the qualitative; (2) the computational stylometric; and (3) the computational linguistic. Within each of these broad approaches are specific formulations of method. Empirical validation of these specific methods rises to different levels, with some specific methods offering more empirical validation than others. Further, some methods demonstrate higher reliability than others, based on empirical validation. Finally, there is a direct relationship, in the United States, between the levels of empirical validation and how authorship evidence can and should be handled as an investigative tool or as one piece of evidence in a trial.

Forensic Stylistics

Forensic stylistics assesses prescriptive errors and "idiosyncrasies" based on the analyst's experience (McMenamin 1993, 2002, Foster 2000, Ollson 2004). The analyst is not required to have training in linguistics, and practitioners in the United States include law enforcement, former speechwriters, literature professors, handwriting examiners. In contrastive terms, forensic stylistics is qualitative, prescriptive, analyst-based and practice-based.

A variation of the forensic stylistics method would include quantification of the prescriptive errors and "idiosyncrasies." In fact, McMenamin (2002) suggests that this variant requires databasing, but at this time the databases have not been fully developed. Without the databases, it is not possible to know the base rate –or typical frequencies—of a feature; without the base rate it is not possible to know if the analyst's perception of a feature as "unique" or "extremely rare" is accurate or biased by the analyst's peculiar experience. Forensic stylistics is the only authorship attribution method whose practitioners claim can peculiarly identify one author, to the exclusion of all others, because the aggregate stylemarkers are considered unique (McMenamin 2002 as well as written opinion reports).

Empirically Validating the Forensic Stylistics Method

Koppel and Schler (2000) applied the standard methodology for empirical validation of the forensic stylistics method. For the linguistic feature set, they used 99 'stylemarkers'which included errors and other idiosyncrasies common in the forensic stylistics literature as well as some lexical and part-of-speech features. For the classification algorithm, they used a linear support vector machine and C4.5 (decision tree) analysis. The support vector machine and C4.5 algorithms are generally accepted as very powerful and very sensitive classification procedures. They used ten-fold cross-validation. For data, they used 480 emails, averaging about 200 words each, written by 11 authors over a year. The highest accuracy for author attribution using the forensic stylistics stylemarkers only was 67.2%. The highest accuracy for author attribution when the forensic stylistics stylemarkers were used with the lexical and part-of-speech tags was 72%. Therefore, it should be noted that Koppel and S chler (2000) could not obtain an accuracy higher than 67.2% with two very powerful classification algorithms using the forensic stylistics stylemarkers. This result tells us that for this large set of forensic stylistics stylemarkers, fed into two very powerful classification algorithms, documents were assigned to their correct authors at best only 67.2% of the time.

Contrastive Linguistics

Contrastive linguistics has been brought to bear on authorship through identifying linguistic interference of specific non-native languages (Kniffka 1993, 2000). Kniffka's method is qualitative, descriptive, analyst-based and theory-based. That is, Kniffka–or indeed any analyst using the method—notes the particular interferences from the native (L1) to nonnative (L2) languages. The analyst must have knowledge of L1-L2 interference based on his own experience as well as the strong and decades-long research literature in contrastive linguistics. Kniffka relies on linguistic theory and predictions from models of language that particular interferences from the native (L1) to nonnative (L2) languages are plausible. L1-L2 errors are not prescriptive errors; they are linguistic behaviors conditioned by the different parameter settings of L1 and L2, and thus descriptively and theoretically grounded.

Although laypersons (or 'natural experts' in Kniffka's terms) certainly recognize linguistic interferences as "foreigner language," the professional linguist has the advantage of actually knowing how particular patterns of interference point to particular parameters of the native language. There is a very rich research literature on second language acquisition and linguistic interference which can be used to analyze texts. Any professional linguist knows of, and knows how to access, this research and use it appropriately.

Kniffka (1993, 2000) does not claim that the contrastive linguistics method identifies one author as opposed to all other authors. Rather, Kniffka shows how the contrastive linguistics analysis describes the linguistic repertoire of the author, and this particular linguistic repertoire can narrow down a list of suspects. (See Chaski and Snider, 2006).

Empirically Validating the Contrastive Linguistics Method

There is a very rich research literature on second language acquisition and linguistic interference which can be used to support the empirical reliability of the contrastive linguistics method. Additionally, Koppel, Schler and Zigdon (2005) applied the standard methodology for empirical validation to the contra stive linguistics method. For the linguistic feature set, they used analyzed each essay for 400 standard function words, 200 letter n-grams, 185 error types and 250 rare POS bigrams, so that each essay was represented by a numerical vector of length 1035, consisting of relative document frequencies for each feature. For the classification algorithm, they used a linear support vector machine with ten-fold cross-validation. For data, they used 258 essays from the Russian, Czech, Bulgarian, French and Spanish subcorpora the International Corpus of Learner English, a corpus of university-level nonnative English essays (Granger, Dagneaux, and Meunier 2002). Koppel, Schler and Zigdon (2005:626) report 80.2% accuracy. In other words, 80% of the essays were correct ly assigned to their native language groups based on th ese particular vectors of linguistic features fed into a support vector machine.

A Slight Paradigm Shift

At this point, we are moving into computational approaches which by definition are quantitative and machine-based. Researchers in both computational stylometry and computational linguistics are working in a paradigm in which empirical validation is part-and-parcel of method development. The standard methodology for empirical validation is the framework for research reports. The goal of this research is to determine the best set of linguistic features combined with the best classification algorithms to produce the highest accuracy scores. Researchers in these fields constantly test combinations of linguistic features and algorithms. The primary difference between the computational stylometry research and the computational linguistics research is the linguistic feature sets' relation to linguistic theory.

Computational Stylometrics

The computational stylometric approach is quantitative, 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). The computational-stylometric approach is descriptive, in that the linguistic features are extracted from the text based on what is actually there rather than on a prescriptive standard. Computational stylometrics is typically machine-based, simply because it is easy to program a computer to calculate the linguistic features, and the linguistic features do not rely on the analyst's subjective response to the documents.

Computational stylometrics derives from a layman's model of language, rather than linguistic theory. It underlying assumptions about language are descriptive, but stylometric features are very superficial, from the perspective of linguistic theory, because they are not connected in any way to the major aims of linguistic theory, such as the acquisition of language, universal properties of language or implicit structure of human language. The analyst need not be a linguist, and most practitioners of computational stylometrics are computer scientists, statisticians, computational linguists and literature professors.

Examples of this approach using various statistical procedures include deVel et al. (2001), Baayen et al. (2002), Stamatatos et al. (2001), Tambouratzes et al. (2004), and Diri and Amasyali (2003). In this line of research, the reports by deVel et al. (2001) and Baayen et al (2002) are very important because of the data they use. In computational stylometry, often the data is literary and/or there are huge amounts of it. But deVel et al.'s (2001) used about 50 emails from three authors, for approximately 12,000 words per author, and Baayen et al (2002) used nine essays on three topics from eight authors, for approximately 9,000 words per author. From a forensic perspective such amounts of data may seem spectacular, but they are far more realistic than the hundreds of thousands of words per author in other computational stylometric studies using literature, newspaper articles and political speeches.

Empirically Validating Computational Stylometrics Methods

DeVel et al (2001) used a very wide array of easily computable, stylometric features: number of blank lines/total number of lines, average sentence length, average word length, vocabulary richness (type/token), total number of function words/tokens, function word frequency distribution (122 features), total number of short words/tokens, count of hapax legomena/tokens, count of hapax legomena/types, total number of characters in words/characters in email body, total number of alphabetic characters in words/ characters in email body, total number of upper-case characters in words/ characters in email body, total number of digit characters in words/ characters in email body, total number of white-space characters/ characters in email body, total number of space characters/ characters in email body, total number of space characters/number white-space characters, total number of tab spaces/ characters in email body, total number of tab spaces/number white space characters, total number of punctuations/ characters in email body, word length frequency distribution/tokens (30 features), has a greeting acknowledgment, uses a farewell acknowledgment, contains signature text, number of attachments, position of requoted text within e -mail body, HTML tag frequency distribution/total number of HTML tags (16 features).

For the classification algorithm, they used a polynomial support vector machine with ten-fold cross-validation. For data, they used 158 email posts from four authors with each author writing on four topics (television, food, travel and movies). DeVel et al. (2001) report high overall accuracies across topics for author 1 (98.2, 96.4 and 86.9). For author 2, only two across-topic accuracies are reported and these are widely spread (27.8 and 82.8). For author 3, only two, but high, across-topic accuracies are reported (96.5 and 98.1); for author 4, only one fairly low accuracy is reported (60). When all of these available accuracies for each author per topic are averaged, the overall accuracy is 81%.

Baayen et al. (2002) demonstrated that lexical and punctuation variables, using nine texts per author on two versions of discriminant function analysis and two versions of cross-validation, obtain cross-validated accuracy rates from 49% to 88%. Baayen et al's (2002) experiments used eight "naïve writers," i.e. first- and fourth-year college students who wrote three texts in three genres (fiction, argument and description). The students were specifically asked to write texts of around 1000 words. When only the lexical, function word variables were included and the cross-validation procedure included texts of all three genres, the standard pairwise discriminant function analyses resulted in an overall accuracy rate of 49%. When the cross-validation procedure was modified so that the genre of the holdout text was matched by the validation texts, the overall accuracy rate improved to 79%. Under the same modification to cross -validation, when the standard discriminant function analysis was enhanced by weighting the vectors by the entropy of the words (so that novel words across texts weigh more than redundant words), the overall accuracy increased to 82%. When 8 punctuation mark variables were added to the lexical function word variables, with the modified cross -validation procedure and the entropy-enhanced discriminant function analysis, the overall accuracy for the 28 author-pairs increased to 88%.

Stamatatos et al. (2001) demonstrated that a totally automatized analysis using syntactic and lexical variables obtains an accuracy rate of 74% to87%. The corpus consisted of 30 texts for each of 10 authors, newspaper columnists writing on a range of topics including biology, history, culture, international affairs and philosophy. The texts ranged in word length from less than 500 words to more than 1,500 words. The linguistic variables used in the linear discriminant function analysis included 50 lexical features and 22 syntactic features. The lexical features were the frequencies of the 50 most frequent words in the training texts normalized for text-length. Using these 50 lexical features, the average (or overall) accuracy (or correct classifications) was 74%. The syntactic features include sentences/words, punctuation marks/words, detected versus potential sentence boundaries, information about phrasal chunks for noun, verb, adverb, preposition and conjunction, and information about parsing such as the number of words unanalyzed after a number of passes. Using the 22 syntactic features, the average accuracy was 81%. When the lexical and syntactic features were combined into a 72-feature set, the highest accuracy rate of 87% was obtained. Especially interesting in Stamatatos et al.'s study is that direct syntactic measures improve on the accuracy rate based on the lexical measures.

Tambouratzes et al. (2004) focused on determining authorship within one register (as defined by general topic). Transcripts of speeches delivered in Greek Parliament by five parliament members over the period 1996-2000 were extracted from a record prepared by the Greek Parliament Secretariat. The speeches ranged in length from less than 300 words to more than 5,000 words. For the linear discriminant analysis, several variable sets of 46, 85, and 25 features were used. These sets included both lexical and syntactic variables. Lexical variables consisted of specific words, while syntactic variables includes part-of-speech (POS) tags and morphological inflections. Other variables relating to word and sentence length as well as punctuation and information about parsing such as the number of tokens unidentified by the tagger were also used. A forward stepwise discriminant analysis with 85 variables indicated that only 25 variables were actually used to generate the classification. The 25 variables included both lexical and syntactic information and are classified as negation, lemmata, micro-structural, macro-structural (such as word-length and punctuation), verbal and part-of-speech. The average cross-validated accuracy rate for the five speakers, using the 85-variable set, and texts of any length was 85%.

Computational Linguistics

Generally similar to the computational stylometric paradigm, a computational linguistic approach is quantitative, descriptive and machine-based. But the features used in the linguistic approach are derived from standard linguistic theory and psycholinguistic experimentation rather than the layman's conception of language or the experience of the natural expert. The syntactic analysis method uses standard syntactic analysis from the dominant paradigm in theoretical linguistics over the past forty years and all the features are justifiable in terms of psycholinguistic processing (Chaski 1997, 2001, 2004, 2005). Gamon (2004) has also developed a method drawing on linguistics. Hirst and Feiguina (2006) have just recently reported a method based on a combination of stylometric and linguistic features.

Empirically Validating Computational Linguistics Methods

For the classification algorithm, Gamon (2004) used a linear support vector machine with five-fold cross-validation. For data, Gamon used 5,767 5-sentence chunks of literature from the three Bronte sisters' novels. Gamon's variables --part-of-speech trigrams, semantic features and most frequent words -- form a vector of 1421 features, but with the data that he has available and the fact that a support vector machine can use more variables than data without the overfitting from the "curse of dimensionality," this large number of variables is not a problem. Gamon (2004) reports a maximum classification accuracy of 85%. considering that Gamon's data consists of only five sentences per document, 85% is rather spectacular result and his method certainly has some potential for forensic applicability.

Hirst and Feiguina (2006) also used a support vector machine, with ten-fold cross-validation, for the classification algorithm. For data, Hirst and Feiguina used 2,232 approximately 200 word segments from the novels of two of the Bronte sisters. Hirst and Feiguina's variables combine stylometric features such as sentence length, function word and punctuation frequencies with the linguistically-motivated part-of-speech bigrams. They report an overall accuracy of 92.4%. Again, given that the texts are so short – approximately 200 words—at an accuracy over 90%, Hirst and Feiguina's (2006) method certainly has some potential for forensic applicability.

Chaski's (2001, 2005) features include syntactic markedness patterns, syntactically-classified punctuation and word length. While word length is a standard stylometric feature, the other features are purely motivated by standard syntactic analysis and psycholinguistic processing.

First, the reality of constituent structure is a well-founded psycholinguistic fact. In the famous "click" experiments, a dichotic listening paradigm is employed. Subjects listen to sentences and a clicks which are placed at different syntactic positions. Subjects are asked to indicate the placement of the click. The very robust result of this design is that subjects place the click at the edge of syntactic structures. For instance, even if the click is positioned one word inside a syntactic structure, the subject will "hear" the click as though it is actually at the edge, or one word away from its auditory position. The click experiments demonstrate the reality of syntactic edges (see Fodor and Bever 1965, and discussion in Garman 1990). One feature of minimalist syntax is the recognition that edges have a bounding effect.

One property of punctuation is that it also marks syntactic edges. Periods, for instance, mark the edges or ends of sentences; commas can mark the edge or ends of clauses or phrases. Chaski (2001) in fact showed that punctuation classified by the type of syntactic edge being marked (sentential, clausal, phrasal or morphemic) was a reliable indicator of authorship.

Second, in many experiments with various designs and tasks, it have been shown that humans are better able to process simple syntactic structures than complex syntactic structures. Simple syntactic structures are easy to process because the internal relations within the structures are easy to resolve through attachment, but complex structures are more difficult to process because of ambiguity or the possibility of multiple attachments, either internal or external to the structure. Probably the most famous example of processing complexity is the sentence "The horse raced past the barn fell." The word "raced" is usually parsed as a finite verb, but that attachment becomes problematic when the actual finite verb "fell" appears. Then "raced" must be reassigned, now attached as a participle to "horse." In terms of creating the subject noun phrase, the first simple processing takes the subject noun phrase as "the horse," while the second complex processing (and the one actually required by this sentence) takes the subject noun phrase as "the horse raced past the barn" (or more clearly as "the horse which was raced past the barn"). For discussion, see Garman (1990) and references therein.

One property of all languages is markedness, the asymmetry which enables a binary contrast to be unequal in complexity and frequency. Syntactic structures can be either marked (complex to process, not as frequent, complicated internally) or unmarked (simple to process, frequent, with a few simple internal relations). It makes sense, psycholinguistically, to hypothesize that markedness in syntactic structures may be indicative of authorship, and, in fact, Chaski (1997, 2001, 2005) has shown empirical support for this hypothesis.

Third, word length and frequency have been known to be correlated in language for a long time (Zipf 1935). Psycholinguistically, also, frequency has been shown to relate to speed of processing, such that highly frequent words are processed more quickly than less frequent or rare words. For discussion of word frequency and lexical access, see Garman (1990) and references therein.

In all languages, words vary by length acoustically and in alphabetic or syllabic writing systems, words also vary by length visually. Again, it makes sense, psycholinguistically, to select word length as a variable for authorship, since the reader's processing of two authors who use very different word lengths would also be different.

Empirical validation requires a carefully designed, known dataset. 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). The five women and five men are all white adults who have completed high school 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 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. The authors all wrote on similar topics, across several different genres (narratives, business letters, personal letters).

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 exceeded the words target. Most authors were represented by seven documents to reach the targets of 100 sentences and/or 2,000 words. Table 1 lists the specifics.

Race, Gender	**Topics by Task ID**	**Author ID Number**	**Number of Texts**	**Number of Sentences**	**Number of Words**	**Average in Swordsman, 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 1: Authors and Texts

Please note that subject 16 obviously has longer sentences than the other authors, since she produced ~2,700 words to get to 107 sentences, while author 23 has much shorter sentences than the other authors, since she produced 134 sentences to get to ~2,100 words.

Each text was processed using ALIAS, a program developed by Chaski (1997, 2001) for the purpose of databasing texts, lemmatizing, lexical frequency ranking, lexical, sentential and text lengths, punctuation-edge counting, POS -tagging, n-graph and n-gram sorting, 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, syntax and lexical.

There are three syntactically-classified punctuation variables, two syntactic markedness variables and one lexical variable, for a total of only six variables:

EOC (end-of-clause), EOP (end-of-phrase) and MPH (morphemic)
mmXP, muXP
AWL (average word length).

Given 10 authors, there were 45 pairwise tests of each author paired with each other author (10*9/2 = 45). SPSS version 13 (Statistical Package for the Social Sciences) was used to run linear discriminant function analysis, based on Mahalanobis distance, with leave-one-out cross-validation. When the variable set included syntactically-classified punctuation, phrase markedness and average word length –measured at the document level– and the DFA was run stepwise, using Mahalanobis distance and the default settings for F to enter and F to remove, with prior probabilities based on group sizes, only one author pair had no variables qualify for the analysis under these settings. Table 2 shows the overall accuracy rate is 95%, with individual authors' accuracy rates ranging from 92% to 98%. In Table 2, "nvq" indicates "no variables qualified" 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	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 2: Cross-Validation Accuracy Scores for Document Level Test (LDA)

In a variation on this empirical validation, 100 sentences from each author were chunked into 5-sentence chunks, so that each author was represented by 20 documents.

The feature set included the previously-used variables (marked and unmarked phrase types, syntactically-classified punctuation, average word length) and average sentence length. Several different algorithms classification algorithms were used. Discriminant function analysis with leave-one-out cross-validation was performed using SPSS, while DTREG was used to perform leave-one-out cross-validated logistic regression (LR), ten-fold cross-validated support vector machine with radial basis function (SVM1) and polynominal (SVM2) kernels and decision tree forest. Table 3 shows each author's overall accuracy score.

AUTHOR	LR	LDA	SVM 1	SVM 2	Decision Tree Forest
16	91.67	95.28	95.28	95.83	92.50
23	85.28	88.33	85.28	80.56	82.78
80	80	85.56	80.56	77.78	81.39
90	72.22	76.67	81.94	80.28	76.11
91	81.39	74.17	81.94	78.89	75.28
96	79.17	83.61	78.61	76.11	76.11
97	85.56	85.83	83.61	79.44	82.22
98	79.72	84.17	79.44	76.11	82.50
99	70	82.50	78.06	74.17	80.83
168	83.33	88.89	76.39	79.72	86.39

Table 3: Cross-Validation Accuracy Score for 5-Sentence Chunk Level Test

Admissibility

In the United States, the forensic stylistics approach has been restricted to limited admissibility in both U.S. Federal and Frye State courts and it has been excluded from testimony completely in at least two evidence hearings (U.S. v. van Wyk, New Jersey 2000; Hargett v. Hargett, Placer County, CA 2005; California v. Flinner, San Diego, CA 2003; Beckman Coulter v. Dovatron/Flextronics, Santa Monica, CA 2003).

On the other hand, Chaski's computational linguistic approach has been allowed into trial with full admissibility in Federal court, after a Daubert hearing (Green v. Dalton/U.S. Navy, District of Columbia, 2000) and the computational linguistic approach supplemented by the contrastive linguistic approach has been allowed into trial with full admissibility in a Frye State court (Zarolia v. Osborne/Buffalo Environmental Corp, Annapolis, MD, 1998).

The fact that the forensic stylistics method lacks empirical validation which demonstrates high reliability was crucial in the Courts' decision to exclude or limit such testimony. On the other hand, authorship attribution methods based on computational and contrastive linguistics were admitted without any restrictions on the testimony even though the results were described as experimental and the methods were described as subject to on-going research. The Courts did not require perfect results. The Courts were satisfied, as any normal scientist is, that the methods had been subjected to some empirical validation, independent of litigation, showing that the methods are fairly reliable under certain conditions. With all the differences between science and law, and certainly all the well-documented and beautifully-expressed miscommunications between lawyers and linguists (Kniffka 1994), it is surely a relief to know that, at the end of the day, we all really want the same thing: reliability in authorship attribution.

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.

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., (2005). "Who's At the Keyboard? Authorship Attribution in Digital Investigations." *International Journal of Digital Evidence* 4(2). Spring 2005. www.ijde.org

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.

Fodor, J.A. and Bever, T.G. (1965). "The Psychological Reality of Linguistic Segments." Journal of verbal Learning and Verbal Behavior 4: 414-20.

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

Gamon, Michael. (2004). "Linguistic Correlates of Style: Authorship Attribution with Deep Linguistic Analysis Features." Available at www.acm.org. Digital Library.

Garman, M. (1990). *Psycholinguistics.* Cambridge: Cambridge University Press.

Granger S., Dagneaux E. and Meunier F. (2002) *the International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain (162 pp.).

Hirst, G. and Feiguina, O. "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts." Manuscript, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4.

Kniffka, Hannes. (1989). "Towards a Methodology of a 'Forensic Linguistics': Postulates and perspectives." Paper read at the First Meeting of the International Association of Forensic Phonetics. York, England.

Kniffka, Hannes. (1994). "Understanding Misunderstandings in Court: "La Serva Padrona" Phenomena and Other Mis-communications in Forensic Interaction." *Expert Evidence: The International Digest of Human Behavior, Science and Law.* SLE Publications. Pp. 164-175.

Kniffka, Hannes. (1993). "Towards a Methodology of Forensic Linguistics: Postulates and Perspectives." *Kriminalistik und forensiche Wissenschaft.* Heft 81, p.67-82.

Kniffka, Hannes. (1999). "Anonymous Author Analysis without Comparison Data? A Case Study with Methodological Implications." Paper read at the 4th Biennial International Association of Forensic Linguists Meeting, University of Birmingham, England.

Kniffka, Hannes. (2000). "Anonymous Authorship Analysis without Comparison Data? A Case Study with Methodological Impact." *Linguistische Berichte* 182, p.179-198.

McMenamin, G. R. (2002*). 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.

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.

Zipf, G.K. (1935). *The psychobiology of language.* Boston: Houghton -Mifflin.