Chapter 9

THE KEYBOARD DILEMMA AND AUTHORSHIP IDENTIFICATION

Carole Chaski

Abstract The keyboard dilemma is the problem of identifying the authorship of a document that was produced by a computer to which multiple users had access. This paper describes a systematic methodology for authorship identification. Validation testing of the methodology demonstrated 95% cross validated accuracy in identifying documents from ten authors and 85% cross validated accuracy in identifying five-sentence chunks from ten authors.

Keywords: Forensic linguistics, authorship identification

1. Introduction

The "keyboard dilemma" is a fundamental problem in forensic linguistic and digital forensic investigations. The keyboard dilemma is posed as follows: Even if a document can be traced to a particular computer and/or IP address, how can we identify who was actually at the keyboard composing the document? It is a particular problem in environments where multiple users may have access to the same computer or when users do not have to authenticate themselves to access a particular account. The keyboard dilemma enables defense lawyers to use the "keyboard defense," suggesting that an unknown person obtained access to the suspect's computer. This is a very possible scenario in corporate cubicles where machines are left running and unattended, or in an open access area such as a public library.

This paper focuses on authorship identification as a means for resolving the keyboard dilemma. It discusses the main issues underlying authorship identification, and proposes a syntactic analysis methodology for authorship identification. This systematic methodology for author-

ship identification is implemented in the Automated Linguistic Identification and Assessment System (ALIAS) [1, 3].

2. Authorship Identification

Several cases involving the successful collaboration of digital investigations and authorship identification have been documented [1, 3]. The cases include a suicide note left on a home computer, racially-targeted e-mails to a supervisor sent from an open access machine at a government agency, and an electronic diary maintained in a military research laboratory that was accessible by military and civilian personnel [3]. Other recent cases have involved common access environments such as classrooms, public libraries, Internet chat rooms, news groups and blogs. The discussion of authorship identification begins with a case that involved the author of this paper as a defense expert.

The author was contacted by the defense attorney of a thirty-year-old female high school teacher who was accused of having sexual relations with a seventeen-year-old male student. The prosecution's evidence was based on several love notes found on a classroom computer. The unsigned love notes were purported to have been written by the teacher and the student. The authorship of some of the love notes was fairly obvious based on their content (e.g., if a note mentioned having to take a test, it was assumed to have originated from the student). For a few notes, it was difficult to tell whether the author was the teacher or the student. Based on their content, sixteen notes, comprising 1,749 words, were first classed as the teacher's love notes; seven notes, comprising 470 words, were classed as the student's love notes.

The defendant maintained that she had not authored any of the attributed love notes. She claimed that the computer had belonged to a former teacher who was involved with students, and who had subsequently left the state. Another possible scenario was that the student authored all the love notes, fabricating the entire correspondence.

The defense attorney supplied the author of this paper with copies of all the love notes as well as several e-mail messages that were known to be from the defendant. The defendant's seventeen known writing samples (1,988 words in total) included e-mail correspondence to parents regarding students' progress reports, and e-mail messages to friends and family members. The author of this paper also requested writing samples from the teacher whom the defendant claimed had written the notes, but the defense attorney could not supply the samples due to jurisdictional and constitutional obstacles.

The issues for this case were: Did the defendant write the teacher's love notes? In other words, were the teacher's purported love notes identifiable with or different from the teacher's known e-mails? Alternatively, did the student write the teacher's and student's love notes? Specifically, were the teacher's purported love notes identifiable with or different from the student's purported love notes?

To answer these questions, the author of this paper applied an authorship identification methodology (described later) called ALIAS [1, 3]. The ALIAS methodology showed that the writing samples from the student and the defendant were clearly different, with the two sets of documents being separated with 100% cross validated accuracy. All the defendant's texts were classified correctly, as were the student's purported love notes. Since the student claimed that he had written love notes to the teacher, and the defendant acknowledged her own writing samples, the 100% separation of the student and defendant demonstrated that the ALIAS methodology could be accurately applied to the data at hand. If the ALIAS methodology had separated the two known authors with lower (say 70%) accuracy, further testing would not have been conducted because the authorship prediction model would not have been sufficiently accurate.

Next, the author of this paper used a discriminant function to predict the authorship of the sixteen love notes purported to have been written by the teacher. Of the sixteen notes, only three were classified as the student's writing while the remaining thirteen were classified as the defendant's writing. At that point it became clear that the teacher's claims were not entirely true.

Even though the analysis pointed to the defendant's authorship of the love notes, the notes themselves revealed that the relationship might not have been consummated. The defense attorney was advised that the authorship analysis was not helpful to his client, but that the content of the notes could provide a defense against the charge of having sexual relations with a minor. Even though the defendant continued to insist that she did not author any love notes, she was found guilty of taking indecent liberties with a minor. She was found not guilty of the more serous crime of having sexual relations with a minor.

3. Forensic Authorship Identification

This section describes factors relevant to real-world authorship identification, namely scarcity, mixed text type and brevity. In addition, it discusses a cross validation approach for verifying the accuracy of authorship attribution methodologies.

3.1 Scarcity, Mixed Text Type and Brevity

Authorship attribution is a pattern recognition problem whose goal is to estimate how similar two sets of patterns are from each other, based on patterns of linguistic behavior in documents of known and unknown authorship. The primary challenges to authorship attribution are text scarcity, mixed text type and brevity of questioned and known texts.

Forensically-relevant texts are typically scarce for several reasons. First, investigators are often expected to act as soon as possible to protect victims. If a person has received threatening communications, the police certainly would not tell the victim to wait until thirty more threats are received. Second, it may not be in the best interest of an investigation to make the suspect aware that authenticated (known) samples of his/her writing are being collected. Third, in some scenarios, e.g., suicide investigations, investigators cannot obtain additional writing samples and so the only available comparison documents are what is at hand. Fourth, from the operational security perspective, minimal communication should be used to gather intelligence so that plots can be foiled before they are set into motion.

Related to the scarcity of texts is another aspect of forensic author identification – text type – which non-forensic author identification methods typically never consider. Text type or register of the comparative known documents is often not the same as the register of the questioned document. For instance, if the suspect document is a suicide note, it is rare that the alleged author would have written other suicide notes that could be compared with the questioned document. If the suspect document is a threat and the known documents are also threats, there is no need to determine the suspect document since the known threats already settle the fact that the author sent threats. If the document is a business e-mail, it might have to be compared to blog posts or love letters or corporate reports. Thus, forensic authorship identification usually has to deal with cross-register or mixed text type data.

Many of the newer authorship identification methods derived from machine learning focus on e-mail or blog text. E-mail and blogs are attractive because of their availability and cultural pervasiveness, but their use in validating authorship identification methods should take into account certain caveats. One problem with this kind of data is that e-mails and blog posts cannot be independently authenticated: the researcher trusts that the screen name was used by one person (again, the keyboard dilemma). Since the very issue to be tested is how accurately a method discriminates authors' writings, it is unwise to test the method on data for which authorship is unknown. For instance, in one recent

test [4], it was argued that two writers must have used the same screen name because the test results split the documents from one screen name into two separate classes. But the other obvious interpretation is that the authorship identification method simply erred in the classification. Because the data was not independently authenticated, there really is no way to assess the accuracy of the method.

E-mail and blog data are not ideal testbeds for authorship identification because the documents typically only include one text type on one topic. Ideally, forensic authorship identification methods should accurately discriminate between authors even when the known and comparative texts are different text types on different topics. A forensic authorship identification method is called upon to compare, for instance, a threat letter to business e-mails, listserv posts, love letters, perhaps even a blog narrative, so the method must be able to work across registers and should not be confined to test data of only one text type.

Forensically-relevant texts are typically brief. Threat letters, suicide notes, ransom demands and phony letters of recommendation are text types that, in general, do not lend themselves to verbosity. The Unabomber's Manifesto is obviously an exception.

Given these characteristics of forensic data, forensic authorship identification methods must be able to cope with a minimal amount of brief documents ranging across different text types. If a method cannot function under these conditions, it may neither be feasible nor reliable enough to be used in a forensic investigation.

3.2 Cross Validation

The accuracy of a classification algorithm is tested using a cross validation technique. This involves withholding a portion of the original data from model building and then classifying it using the model. "Leave-one-out cross validation" involves withholding each text from model building and subsequently using the model to predict its class. In contrast, in "n-fold cross validation," a certain fraction, e.g., one-tenth (n = 10), of the data is withheld from model building and later classified by the model. If ten documents are available and the leave-one-out cross validation methodology is employed, then ten models are built using nine documents each and each document is classified using a model it did not help build. The numbers of hits and misses for each model are averaged to obtain the final cross validated accuracy score.

Suppose Authors A and B have provided the same number of text samples, then any text has an equal (50%) chance of being classified correctly as having been written by A or B. If a classification algorithm

based on certain linguistic features returns a cross validated accuracy score of 100%, then it is clear that A and B can be classified correctly much higher than the base rate of 50%. However, if the algorithm returns an accuracy score of 50%, then A and B can only be classified at the base rate or chance level of 50%.

Cross validation accuracy scores answer two questions. First, whether the selected linguistic features can distinguish two authors using the particular classification algorithm. Second, whether the classification algorithm can distinguish the two authors with the particular set of linguistic features.

Note that an accuracy score does not provide data on the likelihood that Author A wrote a particular text. Rather, the accuracy score is a record of hits and misses for the classification algorithm.

4. ALIAS Methodology

ALIAS (Automated Linguistic Identification and Assessment System) is a syntactic forensic authorship identification methodology that was developed specifically to cope with scarcity, brevity and mixed types [1, 3]. ALIAS has been validated using the Writer Sample Database [1, 3], a forensically-realistic and linguistically-controlled testbed. This section describes the ALIAS implementation, the Writer Sample Database and the validation testing results.

4.1 ALIAS Implementation

ALIAS was developed by Chaski [1, 3] for the purpose of storing, accessing and analyzing texts in a forensically-relevant manner. ALIAS combines data management and computational linguistics tools. Searching and sorting are automatically optimized by building the system within a database platform; natural language analysis is implemented via scripts. ALIAS is built on the Filemaker Pro platform, which provides rapid development and testing on Windows and Macintosh platforms as well as a robust scripting language and plug-ins.

ALIAS includes routines for analyzing natural language and for calculating results based on these analyses. ALIAS implements numerous standard methods and algorithms, including tokenizing, lemmatizing, stemming, n-graphing, n-gramming, sentence-splitting, text-splitting, punctuation tagging, part-of-speech tagging, abbreviation-tagging and phrasal parsing. While ALIAS implements all of these routines for the English language, it can handle multilingual data and non-Roman orthographies for many of these routines.

Tokenizing breaks each text into its words or tokens. Lemmatizing or typing produces the base or dictionary form of a word, from which all other variants of the word can be derived. Stemming, which often overlaps with lemmatizing, produces the main stem by stripping away prefixes and suffixes, e.g., stemming produces "book" from "rebooked" by stripping the prefix "re" and suffix "ed." The n-graphing routine breaks text into a substring of characters or graphs of a length specified by n; n-gramming analyzes the text into a substring of words (or tokens or lemmata or stems or part-of-speech tags) of a length specified by n. Sentence-splitting breaks text into its component sentences so that sentences and structures within the sentences can be analyzed. Text-splitting breaks text into a specified number of sentences. Punctuation-tagging identifies each punctuation mark in the text, while part-of-speech tagging identifies for each word its specific grammatical function (such as noun, adjective, modal verb, finite verb, etc.). Abbreviation-tagging identifies common abbreviations. Phrasal parsing gathers part-of-speech tags into phrases which the words form.

Given the outputs of these routines, ALIAS performs many important computational linguistic calculations. These include calculating type-token ratios for words, lemmata and stems; frequency counts for words, lemmata, stems, punctuation marks, POS tags, n-graphs, n-grams, POS n-grams, sentences, abbreviations, paragraphs within texts, and user-specified patterns; and the lengths and average lengths of words, lemmata, stems, sentences, paragraphs and texts. Additionally, ALIAS produces frequency counts of proprietary patterns related to the distribution of punctuation and syntactic phrase types.

ALIAS offers several functions that are very useful to investigators: authorship identification, intertextuality, threat assessment, interrogation probing and dialectal profiling. Each function implements linguistic methods that have been validated independent of any litigation. ALIAS enables investigators to answer important questions such as: Who authored a text? How similar are two texts? Is this text more like a real threat or a simulated threat? Are there indications of deception in this text? What demographics are associated with this text?

Each function uses specific combinations of the routines described above to produce linguistic variables that are analyzed statistically. The analysis uses simple statistical procedures within ALIAS, but ALIAS also interfaces with SPSS and DTREG [5] statistical software. Other statistical programs such as SAS or Weka [6] may also be used.

Within the authorship identification component, ALIAS includes routines for lemmatizing, lexical frequency ranking, calculating lexical, sentential and text lengths, punctuation-edge counting, POS-tagging, n-

graph and n-gram sorting, and markedness subcategorizing, all based on standard linguistic theory and computational linguistic algorithms. ALIAS is thus able to incorporate a large number of linguistic variables.

Authorship identification in ALIAS uses only three types of variables: punctuation related to syntactic edges, syntactic structures sorted by internal structures and word length. ALIAS produces numerical outputs for these variables that are analyzed statistically. SPSS is used for discriminant function analysis, and DTREG is used for support vector machines and tree model computations.

4.2 Writer Sample Database

The Writer Sample Database [1, 3] was created for the purpose of empirically validating forensic linguistic methods. The database incorporates a collection of texts from 166 subjects of both genders and several races. For inclusion in the study, subjects had to: (i) be willing to write between three to ten texts at their leisure for a small payment; (ii) be students or be employed in positions for which writing was a part of their normal lifestyle; (iii) be similar to each other in dialectal backgrounds so that the methods could be tested to discern sub-dialectal level differences as well as dialectal similarities; and (iv) have at least a senior-year high school educational level and be at least seventeen years of age.

Several factors affected the selection of topics for writing samples. Research has shown that the social context and communicative goal of a message affects its form. It is also known that intra-writer performance varies when the writer is writing for home or personal consumption as opposed to business or professional purposes. Yet, as discussed earlier, forensic methods must accommodate comparisons between two disparate types of text. The tasks also have to be similar to the kinds of text that are actually forensically relevant. To evoke both home and professional varieties, and emotionally-charged and formal language across several text types, subjects were asked to write about the following topics.

- Describe a traumatic or terrifying event in your life and how you overcame it.
- Describe one or more persons who have influenced you.
- What are your career goals and why?
- Write a letter of complaint about a product or service.
- Write a letter to your insurance company.
- Write a letter of apology to your best friend.

Table 1. Demographics of subjects in validation testing data.

Subject	Race	Sex	Age	Education
16	White	Female	40	College (1)
23	White	Female	20	College (2)
80	White	Female	48	College (3)
96	White	Female	39	College (3)
98	White	Female	25	College (2)
90	White	Male	26	College (1)
91	White	Male	42	College (3)
97	White	Male	31	College (3)
99	White	Male	22	College (4)
166	White	Male	17	High School (4)

- Write a letter to your sweetheart expressing your feelings.
- What makes you really angry?
- Write an angry or threatening letter to someone you know who has hurt you.
- Write an angry or threatening letter to a public official or celebrity.

4.3 Validation Testing Data

Ten subjects were selected for the validation tests. Table 1 presents demographic data pertaining to these subjects.

Each subject was set a target of at least 100 sentences and/or approximately 2,000 words, sufficient data to produce reliable statistical indicators. One author (Subject 98) needed only four documents to hit both targets because she produced numerous long sentences. Two authors (Subjects 80 and 96) needed ten documents to produce at least 100 sentences and/or 2,000 words. Three authors (Subjects 16, 91 and 97) needed six documents to hit the sentence target, but only one of the three exceeded the word target. Subject 16 wrote very long sentences that produced the largest number of words, although she produced only 107 sentences in six documents. Details about each author's data are shown in Tables 2 (Females) and 3 (Males).

4.4 Document Level Testing

In many forensic situations, the issue is whether or not a particular document has been authored by a particular person. Therefore, the first validation test was run at the document level.

Table 2. Female authors and texts in the validation testing data.

Subject	Task ID	Texts	Sentences	Words	Av. Test Size (min, max)
16	1-4, 7, 8	6	107	2,706	430 (344, 557)
23	1-5	5	134	2,175	435 (367, 500)
80	1-10	10	118	1,959	195 (90, 323)
96	1-10	10	108	1,928	192 (99, 258)
98	1-3, 10	4	103	2,176	543 (450, 608)
Total		35	570	10,944	

Table 3. Male authors and texts in the validation testing data.

Subject	Task ID	Texts	Sentences	Words	Av. Test Size (min, max)
90	1-8	8	106	1,690	211 (168, 331)
91	1-6	6	108	1,798	299 (196, 331)
97	1-7	6	114	1,487	248 (219, 341)
99	1-7	7	105	2,079	297 (151, 433)
166	1-7	7	108	1,958	278 (248, 320)
Total		34	541	9,012	

ALIAS extracted linguistic patterns (punctuation related to syntactic edges, syntactic structures sorted by internal structures and word length) for each sentence of each document. Next, the sentence output for each document was totaled to obtain the document level data. For each document, the document level counts were divided by the total number of words in the document. This normalization procedure regulated the differences in document lengths.

The document level data was then analyzed using SPSS's linear discriminant function analysis (LDFA) procedure. The procedure was set to use leave-one-out cross validation. Every author was tested against every other author, resulting in a total of 45 tests.

Table 4 presents the results for pairwise testing. For example, the first column of Table 4 shows that Subject 16's documents were discriminated with 100% cross validated accuracy from Subject 23's documents, Subject 80's documents, etc., but were discriminated with only 80% cross validated accuracy from Subject 98's documents. The author average shows the average of the cross validated accuracy scores for an author against all other authors in the testbed. Subject 16's documents on av-

Table 4. Cross validation accuracy scores for document level testing.

Subject	**16**	**23**	**80**	**90**	**91**	**96**	**97**	**98**	**99**	**166**
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
166	100	100	100	80	100	100	100	100	93	X
Average	97	98	94	92	95	93	98	94	92	97

erage are accurately distinguished 97% of the time from the documents of the other nine subjects. Subjects were not tested against themselves, so the corresponding test cells are marked with an "X." Test cells are marked with an "nvq" when no variables qualified for the linear discriminant function.

Table 4 shows that the ALIAS methodology provides an overall accuracy of 95% for the validation testbed.

4.5 Five-Sentence-Chunk Level Testing

In some cases no more than 100 sentences may be available for analysis. The sentences could come from one document or from several short texts. The investigator might have only been able to obtain a sentence or two from one place and a few sentences from somewhere else. The primary issue is the level of reliability that the ALIAS methodology can provide for extremely short documents. This issue is addressed by running a validation test with sentence level data bundled into five-sentence chunks.

As explained earlier, ALIAS outputs sentence level data. Therefore, the first 100 sentences from each of the subjects were used in the test so that every author was represented by 100 sentences. The 100 sentences came from different documents, and a five-sentence chunk might come from two different documents.

SPSS's linear discriminant function analysis (LDFA) procedure and machine learning classification algorithms from DTREG were used in the experiment, along with the leave-one-out cross validation strategy. Ten-fold cross validation was used for the support vector machines with

Table 5. Results for five-sentence chunk level testing.

Subject	Documents	SVM	RBF	Polynomial	DT Forest
16	97	95.28	95.28	95.83	92.50
23	98	88.33	85.28	80.56	82.78
80	94	85.56	80.56	77.78	81.39
90	92	76.67	81.94	80.28	76.11
91	95	74.17	81.94	78.89	75.28
96	93	83.61	78.61	76.11	76.11
97	98	85.83	83.61	79.44	82.22
98	94	84.17	79.44	76.11	82.50
99	92	82.50	78.06	74.17	80.83
168	97	88.89	76.39	79.72	86.39
Average	95	84.50	82.11	79.89	81.61

radial basis function (RBF) and polynomial kernels. The default out-of-bag validation provided by DTREG was used for the decision tree forest (DT Forest) predictive model.

The intent is to show how the different algorithms perform on the data. Therefore, Table 5 does not report the pairwise results for each author; instead, it presents the average for each author and the author average from Table 4 to contrast the document level data and the sentence-chunk level data.

Table 5 shows that Subject 16's five-sentence chunks are highly distinguishable from the nine other authors for an overall average hit rate of 95.28% for discriminant function analysis and support vector machine with radial basis function kernel, 95.83% for the support vector machine with polynomial kernel, and 92.50% for the decision tree forest. Subject 91 shows the lowest average hit rate for discriminant function analysis at 74.17%, with an improved rate of 81.94% for the support vector machine with radial basis function. For the ten subjects, the highest overall accuracy rate of 84.50% was achieved using discriminant function analysis, but the support vector machine and decision tree forest algorithms achieved only slightly lower total hit rates of approximately 80%.

The results for the five-sentence chunk data are lower than the rates for the document level data. But these results are surprisingly good when one considers that the "documents" contained only five sentences. The results suggest that there is at least an investigative function for authorship attribution in the forensic setting, even when the textual data available to investigators is in small five-sentence chunks or in even smaller chunks that are put together to create five-sentence chunks. The

methodology thus appears to hold promise for analyzing blog posts, e-mail messages, pornographic requests and, perhaps, chat room conversations.

5. Conclusions

The ALIAS methodology is a powerful syntactic analysis technique for authorship identification. The validation testing results indicate that the methodology is effective even when the textual data available for analysis is in small chunks, including chunks that are put together from even smaller chunks.

To effectively address the keyboard dilemma, it is important that investigators use a forensic linguistic method that has been validated on forensically-feasible data independently of any litigation. The forensic linguistic method should also have a track record of admissibility. Known writing samples should be authenticated independently and reliably. If samples cannot be authenticated or if there is a possibility that a suspect or attorney may not be telling the truth about the authorship of a known writing sample, it is important to seek out other known samples that can be authenticated.

As in any science, precautions must be taken to avoid confirmation bias. Therefore, the forensic linguist should not be exposed to any details of the case prior to conducting authorship analysis. A digital forensics investigation should be conducted to complement the analysis if the case and the evidence warrant such an investigation. But it is important that the forensic linguist not see the results of the digital forensic analysis and the digital forensics analyst not be privy to any results from the linguistic analysis. The independent convergence of results would serve to strengthen both analyses.

References

[1] C. Chaski, Who wrote it? Steps toward a science of authorship identification, *National Institute of Justice Journal*, vol. 233, pp. 15–22, 1997.

[2] C. Chaski, A Daubert-inspired assessment of language-based author identification, Technical Report NCJ 172234, National Institute of Justice, Washington, DC, 1998.

[3] C. Chaski, Who's at the keyboard? Recent results in authorship attribution, *International Journal of Digital Evidence*, vol. 4(1), 2005.

[4] J. Li, R. Zheng and H. Chen, From fingerprint to writeprint, *Communications of the ACM*, vol. 49(4), pp. 76–82, 2006.

[5] P. Sherrod, DTREG: Software for predictive modeling and forecasting (www.dtreg.com).

[6] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, California, 2005.