Linguistic Authentication and Reliability

Carole E. Chaski, Ph.D.

1.1. Authorship in an Electronic Society

Many different types of crime and civil action involve documents whose origins or authorship must be authenticated. The traditional method of linking document with author has involved Questioned Document Examination, in particular handwriting or typewriter identification and/or ink dating.

But our society is rapidly moving beyond pen, pencil and typewriter; we produce more and more electronic documents. Documents composed on the computer, printed over networks, faxed over telephone lines or simply stored in electronic memory preclude traditional handwriting identification. When the authorship of an electronically produced document is disputed, the analysis of handwriting and typing obviously do not apply, but also in the case of networked printers-- to which thousands of potential users have access --even ink, paper and printer identification cannot narrow the range of suspects or produce a solitary identification. The language of a document, however, is independent of whether a document is written or printed or faxed or stored electronically. The question then arises: can the language of a document be used to link the document with the author?

Since the early 1900's, American courts have dealt with this question, from a legal perspective, in terms of admissibility of language evidence. Table 1 summarizes what has been proffered as language-based evidence of authorship: punctuation, grammatical errors, spelling errors, sentence beginnings, "stylistic deviation."

Date	Case Reference	Type of Linguistic Evidence	Admissible?
1901	Throckmorton v. Holt (1901) 180 US 552, 45 L Ed 663, 21 S Ct 474	punctuation, grammatical errors	*not admissible through expert opinion, but admissible for jury to decide*
1909	State v Kent (1909) 83 Vt 28, 74 A 389	punctuation	yes
1914	Josephs v Briant (1914) 115 Ark 538, 172 SW 1002	spelling, grammatical errors	yes
1916	Bartholomew v Walsh (1916) 191 Mich 252, 157 NW 575	punctuation	yes
1919	Re Fleming's Estate (1919) 265 Pa 399, 109 A 265	spelling	implied admissible
1920	Murphy v Murphy (1920) 144 Ark 429, 222 SW 721	spelling	yes
1929	Re Creger's Estate (1929)	spelling, vocabulary	implied admissible

	135 Okla 77, 274 P 30, 62 ALR 690		
1934	Re Ridley's Will (1934) 151 Misc 474, 273 NYS 48	spelling, grammatical errors	**yes**
1935	State v Hauptmann (1935) 115 NJL 412, 180 A 809, cert den 296 US 649, 80 L Ed 461, 56 S Ct 310	spelling	**yes**
1936	Re Bundy's estate (1936) 153 Or 234, 56 P2d 313	punctuation	**yes**
1943	Re Young's Estate (1943) 347 Pa 457, 32 A 2d 901, 154 ALR 643	signature structure	**yes**
1952	Re Cravens' Estate (1952) 206 Okla 174, 242 P2d 135, 34 ALR2d 615	punctuation	**yes**
1954	Succession of Prejean (1954) 224 La 921, 71 So 2d 328	vocabulary	**implied admissible**
1955	New York v Henry & Armand Mulvey, (1956) conviction rev'd, 1 App. Div .541, 151 N.Y.S.2d 587	sentence length (cf. Menicucci, 1977)	**yes**
1963	Hughes v United States (1963, CA10 NM) 320 F2d 459, cert den 375 US 966, 11 L Ed 2d 415, 84 S Ct 483	spelling	**yes**
1964	Cutler Estate (1964) 33 Pa D & C2d 682	spelling	**yes**
1973	Succession of Killingsworth (1973, La) 292 So 2d 536	vocabulary	**implied admissible**
1976	United States v Pheaster (1976, CA9 Cal(544 F2d 353, 2 Fed Rules Evid Serv 593, cert den 429 US 1099, 51 L Ed 2d 546, 97 S Ct 1118	spelling	**implied admissible**
1976	United States v Hearst (1976, ND Cal) 412 F Supp 893	vulgarity, breathing patterns and pauses, sentence beginnings (cf. Menicucci, 1977)	*no– due to Frye criterion and materiality*
1979	United States v Larson (1979, CA8 Minn) 595 F2d 759	spelling	**yes**
1982	Re estate of Ciaffoni (1982) 498 Pa 267, 446 A2d 225, 36 ALR4th 595, cert den 459 US 1036, 74 L Ed 2d 602, 103 S Ct 447	stylistic deviation	**admissible through expert testimony**
1983	United States v Clifford (1983), CA3 Pa) 704 F2d 86, 12 Fed Rules Evid Serv 870	spelling, punctuation, format, grammatical errors	**District Court-no Court of Appeals-yes**
1984	United States v Campbell (1984, CA1 Mass) 732 F2d 1017	spelling	**implied admissible**
1990–	case rulings do not appear in Lexis-Nexis Search, are self-reported by expert in publication or by personal communication	spelling, punctuation, format, grammatical errors, L1/L2 interference, sentence beginnings	**yes**

Table 1: Summary of Decisions concerning Language-Based Evidence

The judicial record makes two points clear:

2. admissibility is not uniform;

3. the techniques used for determining authorship rely on common misconceptions about language.

Table 1 shows that most of what has been offered as language-based evidence of authorship is exactly the kind of common knowledge which is emphasized in American education: grammatical errors, vocabulary, spelling mistakes, punctuation and style. Further, when the academic and forensic literature is examined, these same ideas come up repeatedly, although they are dressed up in academic jargon. For a technical review of the academic and forensic literature, see Chaski 1998a. Table 2 lists common misconceptions of language use and the academic/forensic techniques which correlate with them.

Common Misconceptions of Language Use	**Techniques**
Individuals have distinct vocabularies.	Type-Token Ratio Hapax Legomena
Individuals use the same words over and over.	Type-Token Ratio Hapax Legomena
Individuals can be identified by the way each says things, i.e. by the *words* each chooses.	Type Token Ratio Hapax Legomena Content Analysis
Individuals can be identified by how sophisticated or simple their sentences are.	Readability Scores Sentence Complexity
Individuals do not share spelling mistakes; spelling mistakes are so rare they can identify users.	Spelling Errors
Individuals do not share grammatical errors; grammatical errors are so rare they can identify users.	Grammatical Errors

Table 2: Common Conceptions of Language Use Related to Techniques

Now the question becomes much more interesting: do the techniques based on common misconceptions about language use actually work reliably and accurately to identify the authors of suspicious documents?

This is a question that can be tested empirically, and my research fellowship at the National Institute of Justice focused on empirically testing methods of language-based author identification.

Before we turn to these results, there is another type of language-based author identification technique based on style and literary interpretation, or literary imagination, which is currently enjoying some notoriety due to the JonBenet Ramsey case.

The New York Times published an interview with Professor Donald Foster about his work as a language expert (Metro Section of City Edition, November 19, 1997). Included with this was his analysis of the ransom note which begins "Listen carefully!" Professor Foster's analysis of these first two works follows.

> The author imagines the text as a heard document, as in a flim kidnapping or a literal dictation (one person speaking , the other writing). A cinematic thread ... includes diction associated with films like "Ransom," "Dirty Harry" and "Speed." A corporate thread ... includes diction associated with a chief executive officer, day-to-day business concerns or computer equipment, possibly indicating a businessperson as author, and/or someone wishing to implicate John Ramsey.

All of this is an interpretation of just the first two words! This is rather impressive, but it is not science. Science, unlike literary criticism, requires that the method of analysis be so clear that anyone who cares to can repeat the analysis and come up with similar results. The method must be objective so that anyone can do it. The method must be quantitative so that the procedure can be standardized. Science is about predictability. But literary criticism, on the hand, strives for originality and dreads replication. What Professor Foster does may be excellent literary criticism, but it cannot be replicated, because it relies on subjective and non-quantitative

interpretation. Therefore, Foster's work, as it is presented in the New York Times interview, cannot generate hypotheses that can be tested empirically.

2.1. Empirical Testing of Nine Hypotheses

There are, however, nine hypotheses for language-based author identification suggested in the literature (for review see Chaski 1998a). Many of these hypotheses have not been replicated in a forensically plausible way because in fact they derive from literary criticism. But it is possible to test these nine hypotheses empirically because they are objective and quantitative. These are:

1: Vocabulary richness identifies authors.
2: Hapax Legomena identifies authors.
3: Readability measures identify authors.
4: Content Analysis identifies/ discriminates between authors.
5: Spelling errors identify authors.
6. Grammatical errors identify authors.
7: Syntactically-classified punctuation discriminates between authors.
8. Sentential complexity identifies authors.
9: Abstract syntactic structures differentiate and identify authors.

2. 2. Empirical Testing of Language-Based Author Identification Techniques

In order to test empirically the current techniques for language-based author identification, a Writing Sample Database was first assembled.[i1] Assembling a database for testing the hypotheses is an essential and time-consuming step which non-scientists are often puzzled by. But in true science, the results are only as good, as reliable, as the experimental design that gets you those results. If there is any question, for instance, as to who actually authored a document, then that document cannot be used experimentally to test

a hypothesis. Therefore great care has been taken to ensure that the Writing Sample Database is designed properly and that data has been collected properly.

A set of four writers was extracted from the database in order to control for sociolinguistic factors which we know affect linguistic performance. This pilot subset mimicks the kind of data which are actually obtained in real casework.

In real casework, the analyst is typically given the unknown, suspect or questioned document(s), and known writing samples from one or more potential suspects. The task is to eliminate some or all of the suspects as the possible author of the questioned document(s) and, if possible, to identify one of the suspects as the possible author of the questioned document(s). In effect, the analyst must distinguish between documents written by different writers and cluster together documents written by the same writer. Both the questioned and known documents are typically short in word length.[ii] Since the investigators have already developed suspects for independent reasons in the typical case, the task of author identification in casework is circumscribed by the number of known sets, and the sociolinguistic characteristics of the known writers such as age, race, sex, and education.

The parameters of real casework have determined the design of the empirical tests. First, the task in all the empirical tests that follow is the same: to distinguish between different writers and to identify documents by the same writer, some known and one unknown, using one particular technique.

Second, the known writing samples were selected on the basis of demographic characteristics which would make the writers similar enough to qualify as a list of suspects. Also, from a theoretical perspective, we know that certain demographic characteristics affect linguistic performance, so a group of people sharing these sociolinguistically significant characteristics would very likely share dialect features. By selecting our "list of suspects" so that they share group or dialect features, we can test a language-based identification technique's ability to go to the individual (or idiolectal) rather than group (or dialectal) level of linguistic performance. Based on both investigative practice and sociolinguistic fact, four writers were selected, from the Writing Sample Database, to form the Pilot Subset. The subject identification numbers and sociolinguistic characteristics of the four writers are shown in Table 4.

Subject ID	Sex	Race	Age	Educational Level	Dialect Information
001	F	Black	40	College 2	US Delmarva
009	F	Black	47	College2	US Delmarva
016	F	White	40	College1	US New England & Delmarva
080	F	White	48	College3	US Delmarva

Table 3: Subjects in the Pilot Subset

Subject ID	Sex	Race	Age	Educational Level	Dialect Information
001	F	Black	40	College 2	US Delmarva
009	F	Black	47	College2	US Delmarva
016	F	White	40	College1	US New England & Delmarva
080	F	White	48	College3	US Delmarva

Table 4: Subjects in the Pilot Subset

Third, as in actual casework, the writing samples from these four subjects are short. The shortest text contains only 93 words, the longest, 556. Three texts were used from subjects 001, 009 and 080, while only two were used from subject 016, in order to keep the number of words from the subjects relatively comparable. In this way, subjects 001 and 080, and subjects 016 and 009, respectively, produced a comparable number of words. Since most questioned documents are short, the goal is to test techniques on short documents. In fact, it is important to develop techniques which can operate successfully on short documents, as the worst case scenario, even if long documents are available in particular cases. The textual characteristics of the Pilot Subset are shown in Table 5.

Table 5 also shows the number of words in the questioned document (QD). The QD text was selected by an intern at the National Institute of Justice from the documents generated by the four writers, typed into the computer, identified as SQD2. The true

identity of SQD2 was not revealed to the analyst until after the empirical tests were conducted. So the analyst knew that the document was authored by one of the four writers but not which one.

Subject ID	Number of Texts Used	Number of Words in Text of Topic 1	Number of Words in Text of Topic 2	Number of Words in Text of Topic 3	Number of Words in Text of Topic 4	Total Number of Words in Texts
001	3	223	121	187		531
009	3	361	265	372		998
016	2	344	556			900
080	3	239	93	103		345
SQD2	1				341	341

Table 5: Text Characteristics of Subjects in Pilot Subset

3.3. Results of Empirically Testing the Nine Hypotheses on the Pilot Subset

HYPOTHESIS 1: Vocabulary richness identifies authors.

Source: See Holmes (1994) [44] for review and references; Baker (1988) [78].

Methodology:

Count number of total words in text; let N = tokens.
Count number of distinct words in text; let V = types.
Calculate TTR and PACE for texts of each writer.
Compare each writer's TTR and PACE to each other's.

Tools: Type-Token Ratio and Pace.

$TTR = V/N$

$PACE = 1/TTR$

Results In Tabular Format:

Subject ID	Texts	TOKENS*	TYPES*	TTR	PACE
s001	3	527	256	0.4858	2.0586
s009	3	998	373	0.3737	2.6756
s016	2	879	347	0.3948	2.5331
s080	3	435	221	0.5080	1.9683
sQD2	1	341	186	0.5455	1.8333

Table 6: Type-Token Ratio and Pace for Each Writer's Texts

*Note: Due to the small sizes of these texts, all texts written by the author were combined in order to count tokens and types. This could be a false move in a forensic setting if the "known" writing samples are not actually all written by the same writer.

Analysis: The TTRs of subjects 009 and 016 are very similar; likewise, the TTRs of subject 001 and 080 are very similar. TTR clusters together texts from four writers into two groups; in each of these groups, texts from different writers are clustered together erroneously.

The unknown writing sample, QD2, has a TTR which is very similar to subjects 080 or 001. QD2 was actually written by subject 016, not subject 080.

If an analyst relied on TTR, he would mistakenly conclude that he was dealing with two known writers --the clusters of 009/016 and 001/080-- rather than four known writers. Further, he would conclude that the questioned document was authored by the erroneous cluster 001/080, rather than the correct conclusion that it was written by subject 016.

Not surprisingly, PACE (which is just a reciprocal of TTR), leads to the same erroneous inferences.

Replication Results: The hypothesis that vocabulary richness identifies authors has failed to be replicated successfully in a forensically-similar test.

HYPOTHESIS 2: Hapax Legomena (a Greek term for "spoken once") identifies authors.

Source: See Holmes (1994) [44] for review and references; cf. Ule (no date) [79].

Methodology:

Count total number of words in text; let N = tokens.

Count number of words occurring once in text; let V1 = types occurring once.

Calculate Ratio of Hapax Legomena to Tokens (HLR) for texts of each writer.

Compare each writer's HLR to each other's.

Tools: Hapax Legomena Token Ratio

HLR = V1/N

Results In Tabular Format:

Subject ID	Texts	TOKENS*	V1*	HLR
s001	3	527	77	0.1461
s009	3	998	213	0.2134
s016	2	879	214	0.2435
s080	3	435	166	0.3816
sqd2	1	341	136	0.3988

Table 7: Hapax-Legomena-Token Ratio for Each Writer's Texts

*Note: Due to the small sizes of these texts, all texts written by the author were combined in order to count tokens and V1. This could be a false move in a forensic setting if the "known" writing samples are not actually all written by the same writer.

Analysis: The HLRs of subjects 009 and 016 are very similar; the HLR of subjects 001 and 080 differ. HLR clusters together texts from four writers into three groups, 001, 009/016, and 080; in one of these

groups, 009/016, texts from different writers are clustered together erroneously.

The unknown writing sample, QD2, has a HLR which is very similar to subject 080. QD2 was actually written by subject 016, not subject 080.

If an analyst relied on HLR, he would mistakenly conclude that he was dealing with three known writers -- 001, the cluster of 009/016 and 080-- rather than four known writers. Further, he would conclude erroneously that the questioned document was authored by 080, rather than the correct conclusion that it was written by subject 016.

Replication Results: The hypothesis that hapax legomena identify authors has failed to be replicated successfully in a forensically-similar test.

HYPOTHESIS 3: Readability measures identify authors.

Sentence length and Word length both factor in most readability measures.

Source: See Ellis and Dick (1996) [55] for an example of this hypothesis; for sentence length and word length see Holmes (1994) [44] for review and references.

Methodology:

Select readability formula.

Apply readability formula manually or by computer (e.g. through word processing programs).

Compare grade level, etc. for each text to other texts.

Tools: Readability formulae, possibly t-test or correlation statistics.

Readability Formula Pilot Test 1 using Pilot Subset

Results in Tabular Format:

Subjects(Texts)	001(3)	009(3)	016(2)	080(3)	QD2(1)
Passive Sentences	9	13	10	5	0

	0	11	41	0	
	9	11		16	
Flesch	74.5	93.1	58.0	80.7	69.9
	56.3	62.5	68.8	73.7	
	71.5	68.5		57.1	
Flesch Grade Level	7.5	5.6	10.5	6.9	8.4
	11.1	8.7	8.1	7.6*	
	7.8	8.1		10.8	
Flesch-Kincaid	7.9	3.3	13.6	5.4	9.0
	10.4	8.1	12.2	6.8*	
	7.5	7.1		9.5*	
Gunning-Fog	10.6	5.8	16.7	8.6	11.3
	14.3	11.7	15.3	8.3*	
	9.1	9.4		14.2	

Table 8: Readability Formulae Results on Pilot Subset

*Note: The Microsoft Word version of these readability formulae reports that the asterisked numbers may not be reliable due to insufficient number of words in the texts.

Analysis: The readability scores for each author's set of documents look similar across all the authors. For instance, the Flesch scores all seem to be in the range of 60 to 70, on the average. Since there is variation among the documents in each author's set, the degree to which each author's texts are similar can first be measured. For this, the correlation statistic is feasible.

The scores for texts written by each subject are, after all, highly correlated; each writer appears to be consistent across different texts in terms of readability scores as shown below:

Correlation Matrices For Readability Scores Within Writers

Subject 001	Texts	01	02	03
	01	1		
	02	.993	1	
	03	1	.992	1
Subject 009	Texts	01	02	03
	01	1		
	02	.993	1	
	03	.997	.999	1
Subject 016	Texts	01	02	
	01	1		
	02	.996	1	
Subject 080	Texts	01	02	03
	01	1		
	02	1	1	
	03	.99	.992	1

One would expect these very high correlations to decrease if the scores from QD2 are added to the wrong writer's scores. But when the QD2 is grouped with each of these different writers, these very high correlations do not decrease, and in fact stay consistently high across the board:

Subject 001		**01**	**02**	**03**	**QD2**
	01	1			
	02	.993	1		
	03	1	.992	1	
	QD2	.999	.996	.999	1
Subject 009		**01**	**02**	**03**	**QD2**
	01	1			
	02	.993	1		
	03	.997	.999	1	
	QD2	.994	1	.999	1
Subject 016		**01**	**02**	**QD2**	
	01	1			
	02	.996	1		
	QD2	.992	.998	1	
Subject 080		**01**	**02**	**03**	**QD2**
	01	1			
	02	1	1		
	03	.99	.992	1	
	QD2	.997	.998	.997	1

If an analyst relied on Readability measures, he might recognize that he was dealing with four known writers, but he would conclude erroneously that the questioned document was authored by any one of these writers, rather than the correct conclusion that it was written by subject 016.

Another way to analyze these data, implemented by Ellis and Dick in their work on Civil War correspondents, is to compare the readability scores of different writers by the t-test. Using the null hypothesis that there is no difference between the readability scores of writers who have previously been clustered by other techniques, consider the t-test results:

writers compared	paired-t value	probability of no difference
001 and 080	-.136	.8986
009 and 016	.211	.8432

001 and 009/016	.335	.7541

What these probabilities tell us is straightforward. Readability scores do not differentiate between writers of similar sociolinguistic characteristics (age, race, sex, educational level and dialect background). It is doubtful however whether readability formulae are even capable of distinguishing between writers who differ on educational and dialect levels. The following data from an actual case included three white men, in their twenties. Two men were Southerners with college degrees. One man was a Northerner with ten weeks to go before receiving his M.D.

Readability Formula Pilot Test 2 (Actual Case Data)

Results in Tabular Format:

Subjects	QD	B	C	D
Passive Sentences	9%	6.8%	4%	5%
Flesch	84.2	80.6	82.5	86.0
Flesch Grade Level	6.5	6.9	6.7	6.3
Flesch-Kincaid	5.7	5.6	4.8	4.8
Gunning-Fog	8.5	8.1	8.3	7.8

Table 9: Readability Formulae Results on Actual Case Data

There is certainly no need for a t-test here! It is obvious that readability scores would never differentiate between the sets of known writers B, C, D or lead to any one of them being eliminated from the authorship of the questioned document.

Replication Results: The hypothesis that readability measures identify authors has failed to be replicated successfully in a forensically-similar test.

HYPOTHESIS 4: Content Analysis identifies/discriminates between authors.

Source: Kenneth Litkowski (personal communication).

Methodology:

Classify each word in document by semantic category.
Analyze statistically the distance between documents.

Tools: Classification scheme based on semantic categories; linear discriminant functions for statistically computing distance between documents.

Professor Donald McTavish ran the analysis of the pilot subset documents and returned an initial report which was forwarded to me by Kenneth Litkowski. Portions of this report are quoted in this summary, but in order to understand them, the reader must understand McTavish's way of labeling the texts, by number and letter, and how these relate to the Pilot Subset ID labels and the thematic topics in each document. These are listed in Table 10.

Pilot ID	Topic	McTavish ID	
001-01	trauma	1	A
001-02	influence	2	B
001-03	goals	3	C
009-01	trauma	4	D
009-02	influence	5	E
009-03	goals	6	F
016-01	trauma	7	G
016-02	influence	8	H
016-03	goals	9	I
080-01	trauma	10	J
080-02	influence	11	K
080-03	goals	12	L
QD2	anger	13	M

Table 10: Correlating McTavish's Identification Scheme with the Pilot Subset IDs

McTavish's Comments on the C-scores, or Context-Scores:

> ...four texts (C,F,K,L) talk about goals, four talk about terror (A,D,I,G), four talk about influential people (B,E,H,J) and one (M) deals with anger. Looking at the 1x2 plot, those talking about goals are on an outer ring, the outliers plus L, which, like C and K, is somewhat more distant on dimension 3. The "terror" texts are generally high Traditional and low Practical. The "influence" texts are lower Traditional and lower Practical but B is an exception (high Traditional). In general there is strong patterning evident. At first I had expected some sort of pairing across the two arcs (B-M-A-E and L-G-D-H-J) but I haven't found the criterion if pairing is going on. ... Overall, there is a pattern in the plots that probably connects with the patterns designed into the data if one knew more about the sources and conditions of the data. The outliers appear to be texts F, K, C and perhaps I.

McTavish's Comments on the E-scores, or Emphasis-Scores:

> I had hoped that theme differentiation would pattern in more obvious ways. It appears that K and J are more positive outliers and M is an outlier in a more negative dislike direction. There is some patterning but it doesn't seem to connect well with discriminating authorship. ...I can suggest that some texts are more different than the others (F, K, C, and perhaps I, contextually; K, J and M conceptually). K seems to be the one that is different in both respects.

Analysis: Semantic categorization of the texts groups together the texts which share the same topics (trauma/terror, influence, goals and anger) through the clustering of Context-Scores. In one "arc" (B-M-A-E) texts from writers 001, QD2 and 009 are clustered, while in another "arc" (L-G-D-H-J) texts from 080, 016 and 009 are clustered. These arcs represent a similarity between 001 and 009, on the one hand, and 080, 016 and 009, on the other. Further, the first arc shows a similarity between the QD2 text and both 001 and 009. The Emphasis-Scores appear to cluster texts from all of the writers (F, K, C, I or 009, 080, 001 and 016) "contextually" and two of the writers (K, J, M or 080 and QD2) "conceptually."

If an analyst relied on Content Analysis's C-scores, he would mistakenly conclude that he was dealing with two known writers: 001/009 on the one hand and 080/016/009 on the other. Further, he would conclude erroneously that the questioned document was

authored by 001/009, rather than the correct conclusion that it was written by subject 016. If an analyst relied on Content Analysis's E-scores, he would mistakenly conclude that he was dealing with two known writers 001/009/080/016 on the one hand and 080 on the other. Further, he would conclude erroneously that the questioned document was authored by 080, rather than the correct conclusion that it was written by subject 016.

McTavish himself recognizes that the semantic categorization of texts is not able to discriminate between authors, when he comments that "there is some patterning but it doesn't seem to connect well with discriminating authorship."

Replication Results: The hypothesis that Content Analysis scores identify authors has failed to be replicated successfully in a forensically-similar test.

HYPOTHESIS 5: Spelling errors identify authors.

Sources: McMenamin 1993 [4]; Janet Randall, Ph.D. (personal communication); Ron Butters , Ph.D. (personal communication).

Methodology:

List each spelling variant in texts of each writer.
Compare spelling patterns.

Tools: Spellcheckers or other dictionaries; knowledge of English spelling patterns

Results in Tabular Format:

Spelling Variants Test 1

Subjects	s001	s009	s016	s080	SQD2
Texts	3	3	2	3	1
Texts w/*sp	2	2	2	0	1
variants:	mos systematicly developement recieve uniquness	wass	structoring nite arguement		espeically

Table 11: Spelling Variants in Pilot Subset

Analysis: Given these lists, 001 and 016 appear to be "poor spellers" while 080 appears to be a "good speller" and 009 is probably a "good speller" who suffered a momentary slip of the pen. 001 texts and 016 texts share one spelling pattern: the [e] before the suffix [ment] in 001's developement and 016's arguement. 001 's uniquness also involves [e] with a suffix but this pattern cannot be related to other patterns outside the 001 set. 001 texts and QD2 text share a mislinearization of the graphemes [c, i, e] in 001's recieve and QD2's espeically. 016 texts and QD2 text show no relation in spelling patterns. Other spelling errors such as 001's systematicly for systematically or mos for months or 016's structoring for structuring and nite for night cannot be related to other patterns in these documents.

If an analyst relied on spelling errors, he would mistakenly conclude that he was dealing with three known writers -- the cluster of 001/016, 009 and 080-- rather than four known writers. Further, he would conclude erroneously that the questioned document was authored by 001 or 001/016, rather than the correct conclusion that it was written by subject 016.

Perhaps the spelling error technique requires more writers in the suspect set. In order to allow for this, another Spelling Errors Pilot Test was conducted. This time the texts written by the first eleven women in the Writer Sample Database were extracted and each spelling error was listed, as shown in Table 12. The first eleven women range in age from 18 to 49, so there is less sociolinguistic control in the second pilot.

Spelling Variants Pilot Test 2

Subject ID	Spelling Variants
001	mos developement systematicly recieve uniquness
002	terifying licences behide realy regestration frount wher
003	somthing wates dispite mostely
004	occured fellas alright
005	tramatic alot differend lattern constitionally beween
006	haveing togather collasped hospal standrads gudided opputunities reaity indivuaal attudute personaily frightnd potental field acheive awre crimal venture'es knowling diffcult
007	recieving
008	--none--
009	wass
010	--none--
011	occuring aroud prepairing prepaired opressed impresionable disfunctional beyound habilatation lifes travisty politicaly racialy

Table 12: Spelling Variants in Expanded Pilot Subset from Writing Sample Database

Analysis: Writers 002 and 011 share several, very similar spelling error patterns.
These are:

1. errors with doubled consonants:

002	terifying	[terrifying]
011	occuring	[occurring]
	opressed	[oppressed]
	impresionable	[impressionable]

2. errors with doubled consonant with suffix [ly]

002	realy	[real + ly > really]
011	politicaly	[political + ly> politically]
	racialy	[racial + ly > racially]

3.errors with vowels preceding nasal consonant

002	behide	[behind]
	frount	[front]
011	aroud	[around]
	beyound	[beyond]

The nasal consonant is dropped in 002's behide for [behind] and 011's aroud for [around]. The vowel preceding the nasal consonant is expanded in 002's frount for [front], and 011's beyound for [beyond].

4. errors with vowel [I] sound as in "sit" [SIT]

002	regestration	[registration]
011	disfunctional	[dysfunctional]
	travisty	[travesty]

These spelling patterns are very similar, but they originate from two different authors. If an analyst relied on spelling errors, he would mistakenly conclude that he was dealing with one known writer -- the cluster of 002/011,-- rather than two known writers.

Likewise, if the common conception of "poor spelling" is used, writers 002, 006 and 011 would be erroneously thought to be one writer, because these three writers are indeed "poor spellers." But these poor spellers are three distinct authors. Similarly, the common conception of "good spelling" would erroneously lead an analyst to conclude that 008 and 010 are one and the same writer because they are both in fact good spellers, but two good spellers, not one.

Finally, the spelling errors technique would be extremely difficult to quantify unless the documents were extremely long and contained repeated instances of spelling patterns. The technique is subjective in that "good" spelling and "poor" spelling can mean different amounts of spelling mistakes to different people. One spelling error may signal "poor speller" to one person on the jury, while five spelling errors may be required to signal "poor speller" to another person on the jury.

The frequency of spelling errors is another issue which should be considered, as Goutsos pointed out with regard to McMenamin's spelling-based analysis. Even errors that appear to me, subjectively, as rare, such as the behide/aroud pattern, are not so odd that they cannot be shared, as shown by writers 002 and 011. Without frequency data it is almost impossible to figure out how to quantify observations based on spelling errors. Linguists who suggest spelling errors as individualistic do not, to my knowledge, quantify their observations, although I believe that McMenamin is considering this.

It is very likely that spelling errors signify group behavior reflective of dialect background, education and auditory processing abilities rather than individuality. Even children who invent their own spellings in preschool activities often follow general rules.

Replication Results: The hypothesis that spelling errors identify authors has failed to be replicated successfully in a forensically-similar test.

HYPOTHESIS 6: Grammatical errors identify authors.

Sources: McMenamin (1993) [4], Janet Randall, Ph.D. (personal communication), Ron Butters, Ph.D. (personal communication)

Methodology: List all grammatical errors in text, using school grammar.

Compare errors.

Tools: Prescriptive grammar books, GrammarChecker in word processing software.

Results in Tabular Format:

	ID Texts	ID Texts	ID Texts	ID Texts	
Subjects	s001 (3)	s009 (3)	s016 (2)	s080 (3)	SQD2
sentence fragment	1 0 2(3)	0 0 0	1 0 (1)	0 0 0	2
run-on sentence	1 0 2(3)	2 0 1(3)	5 0 (5)	0 0 0	2
subject-verb mismatch	0 0 1(1)	0 0 0	0 0	0 0 0	0
tense shift	0 1 0(1)	0 0 0	0 0	0 0 0	0
wrong verb form	0 0 1(1)	0 0 1(1)	0 0	0 0 0	0
missing (aux) verb	0 0 0	2 0 3(5)	0 0	0 0 0	0

Table 13: Frequency of Prescriptive or School Grammar Errors in Pilot Subset
The first three numbers in each column represted the number of the error in the first, second and third text respectively. The number in parentheses is the total number of the error in the writing sample from the author.

Analysis: There are two ways to interpret these data. One is to read the rows or error types as indicative of authorship; the other is to read the columns or error frequency as indicative of authorship.

Reading the rows --or error type-- reveals the following patterns. 001, 009 and 016 all have run-on sentence. 001 and 016 have sentence fragments as well as run-on sentence. 001 and 009 have wrong verb form as well as run-on sentence. 001 has subject-verb mismatch and tense shift which no one else has, separating 001 from 016 in part. 009 has missing auxiliary verb which no one else has, separating 009 from 001 in part.

Thus, if an analyst were dealing with prescriptive grammar errors by error type, he would mistakenly conclude that he had six authors-- a cluster of 001/016, a cluster of 001/009, 001, 009, 016 and the grammatically superior 080. SQD2 could, however, be correctly assigned to 016.

Reading the columns --or error frequency-- reveals the following patterns. 080, 009-02 and 016-02 have no errors. 001 and 009 have the same number of errors (9). 016 has the second most number of errors (6).

Thus, if an analyst were dealing with prescriptive grammar errors by error frequency, he would mistakenly conclude that he had three authors --a cluster of 080/016/009, a cluster of 001/009 and 016. SQD2 could, however, be correctly assigned to 016 on the basis that there are not many errors in the text.

Neither interpretation relies on a statistical test because there are too many zeroes in the frequencies.

It would appear, then, that the grammatical errors technique, if type is used, at least begins to take us to the right answer. It enables us to distinguish between the four writers, and it enables us to cluster the questioned document with the correct writer in the pilot subset, even if it does not enable us to cluster documents from each author correctly.

But this result does not warrant a full-fledged acceptance of the technique for four reasons. First, the whole notion of school grammar, the idea that a native speaker's use of his own language is right or wrong, violates all linguistic theory and descriptive linguistics. There is no defense for this technique having been suggested by academicians trained in modern linguistics except that this is what most people think of when they think of grammar, so it is easy to explain to juries.

Second, since most non-standard dialects are defined in terms of the standard school grammar, it is highly likely that the grammatical errors technique actually confounds class with individual characteristics. As mentioned earlier, handbooks on composition document that there are "ten most frequent errors" (comma splices, it's for its, etc.) found in most non-academic writing (see for instance Berry (1971). So almost by definition grammatical errors belong to groups of people, not individuals.

Third, because prescriptive grammatical errors are so well known and easy to explain, even computers can identify them and in most instances correct them. Word processing programs such as WordPerfect or Word contain grammar checkers which can resolve most of these errors for producers of electronic documents. If person A's known writings contain peculiar errors, but person B's writings are known to be grammatically correct, a clever A might spell-check and grammar-check the fraudulent document. Butters, a forensic linguist, for instance, has mentioned to me his belief that "you can't

perform a rule you don't know." But you can get a computer's word-processing program to perform a rule you don't know. This could lead the error-based analyst to the false conclusion that B authored the document actually composed by A (false identification).

Fourth, the grammatical errors technique is very difficult to quantify. Linguists who have suggested this method do not quantify their results. Partly, this is no doubt because quantifying the errors would involve quantifying the entire document. Suppose, for instance, that errors would be counted as part of a percent of items which includes the number of times the phenomenon was produced correctly. Then all instances of the phenomenon would have to be counted. It is simply much easier not to do this kind of quantification, and it is in fact not even part of the prescriptive grammar tradition to compare rates at which particular "errors" occur (although quantitative sociolinguistics such as Labov's work would require this kind of total quantification).

Fifth, it is possible to keep the baby and throw out the bath water. Analytical techniques based on descriptive linguistics are able to discern the same types of patterns --and more-- without resorting to prescriptive grammar. Further, these same analytical techniques would enable us to quantify the entire document so that rates of particular phenomena could be ascertained.

Replication Results: The Grammatical Errors technique has been partially replicated but is still held in reservation due to theoretical and statistical problems.

HYPOTHESIS 7: Sentential complexity identifies authors.

Source: Svartik (1968) [59].

Methodology: Classify sentences into sentential categories.
Count frequencies of each category.
Test statistically.

Tools: Knowledge of sentential syntactic categories such as simple, compound, complex, and compound-complex or Svartik's own six clausal categories.
Knowledge and use of $\chi 2$ statistic.

Results in Tabular Format:

Subjects (Texts)	s001(3)	s009(3)	s016(2)	s080(3)	SQD2
sentence fragment	1 0 2(3)	0 0 0 (0)	1 0 (1)	0 0 0(0)	2
simple sentence	3 1 5(9)	11 4 16 (31)	3 2 (5)	9 3 2(14)	6
compound sentence	2 0 1(3)	4 3 3 (10)	0 0 (0)	1 0 0(1)	3
complex sentence	5 4 4(13)	11 10 8 (29)	3 3 (6)	4 3 4(11)	7
compound -complex	3 0 1(4)	3 1 0 (4)	5 10 (15)	3 0 0(3)	3
Total sentences:	14 5 13 (32)	29 18 27 (74)	12 15 (27)	17 6 6 (29)	21

Table 14: Frequency Data of Sentence Types in Pilot Subset

Analysis: Svartik's analysis of the confessions in the Timothy Evans case exemplifies both grammatical error analysis as well as the sentential complexity technique. Svartik repeatedly refers to Evans as an "illiterate" who uses "substandard" language. The underlying principle in sentential complexity analysis is the idea that some sentence structures are more complex than others and that people will differ in their abilities to produce different types of sentential complexity.

The hypothesis that patterns of sentential complexity differentiates between writers can be tested statistically, and in fact Svartik used the chi-square test. Assuming the null hypothesis that there is no difference between the sentential complexity patterns of pairs in the pilot subset, what is the chance that these paired patterns come from the same author? The results are shown in Table 15.

	01/09	01/16	01/80	09/16	09/80	16/80
$\chi 2$	28.123	23.76	18.52	52.325	17.152	25.529
p	0.1065	0.0941	0.553	0.0001	0.3099	0.061

Table 15: Statistical Analysis of Sentential Type Data in Pilot Subset Writers

These probabilities suggest that writers 009 and 016 can be clearly differentiated by the sentential complexity method, because the chance of there being no difference between them is so extremely low (1 in 10,000). Further, writers 016 and 080 might be differentiated by the sentential complexity method, because the chance of there being no difference between them is almost acceptable in terms of statistical significance (6 in 100). More disappointing is that the sentential complexity method cannot strongly distinguish between the texts authored by 001 and 009, or 001 and 016, or 009 and 080, or 001 and 080.

The chi-square results in Table 16 relate to the null hypothesis that there is no difference between the punctuation patterns in QD2 and each of the writers in the Pilot Subset. Since the truth is that there is a difference between the author of QD2 and authors 001, 009 and 080, we expect a very low probability of no difference in these pairings, but a high probability of no difference in the pairing of QD2 and 016.

	01/qd2	09/qd2	16/qd2	80/qd2
$\chi 2$	7.209	17.908	13.231	11.122
p	0.8435	0.1185	0.1041	0.5185

Table 16: Statistical Analysis of Sentential Type Data in QD and Pilot Subset

As Table 16 shows, however, these expectations are dashed. Indeed, there is no significant difference between the sentential patterns of QD2 and any of the writers.

If an analyst relied on sentential complexity, he would mistakenly conclude that he was dealing with three known writers -- 009, 016, and a cluster of 001/009/080 texts-- rather than four known writers. Further, he would conclude erroneously that the questioned document was authored by any of these three "authors", rather than the correct conclusion that it was written by subject 016.

Svartik's measure of sentential complexity separated relative clauses from other types of subordinate clauses and counted compound verb phrases as separate clauses. Although this counting may not be completely defensible within generative grammar, it points out that different measuring tools may lead to different results. In fact, measuring really natural language is quite different from measuring edited language or textbook examples. Whenever the measuring device is vague, subjectivity can creep in. Therefore, it is advisable to reserve final judgment on the forensic suitability of sentential complexity as an identification technique until these methodological problems have been resolved.

Replication Results: The hypothesis that sentential complexity patterns identify authors has failed to be replicated successfully in a forensically-similar test; however, this failure to be replicated may be caused by methodological problems in determining how to measure and count sentential complexity.

HYPOTHESIS 8: Syntactically-classified punctuation discriminates between authors.

Sources: McMenamin (1993) [4] suggests that punctuation is idiosyncratic but his approach does not include quantification. Pilot studies presented in an National Institute of Justice Research Seminar, (Chaski 1996) [80], suggested that punctuation which is syntactically classified and subjected to statistical testing may be idiolectal. The methodology which follows comes from Chaski (1996) [80].

Methodology: List each punctuation mark.
Classify by the mark's syntactic function, e.g, End-Of-Sentence period, comma separating main and dependent clauses, comma separating phrase, comma in list, etc.
Test statistically the hypothesis that syntactically-classified punctuation differentiates between writers.

Tools: Knowledge of punctuation and syntax.
Knowledge and use of $\chi 2$ statistic.

Results in Tabular Format:

PUNCTUATION	s001	s009	s016	s080	QD2
EOS .	24	69	21	22	14
EOS . for ?	2	1	0	1	0
EOS no mark	4	1	0	0	0
EOS ?	0	0	1	5	3
EOS !	1	4	0	0	0
()	0	1	2	0	1
"" on S	1	3	8	0	2
"" on W	0	0	3	0	3
' contraction	10	5	15	1	9
' plural	1	0	0	0	0
' possessive	0	2	4	0	2
+ and	1	0	0	0	0
comma in list	0	6	11	3	4
comma main/dep	4	8	12	6	2
comma main/main	1	2	7	0	1
comma for phrase	3	1 2	17	8	2
; in list	1	0	0	0	0
; main/dep	1	0	0	0	0
; main/main	0	1	0	0	0
- between Ss	0	1	41	2	33
: colon	0	0	2	0	0
- in W	0	1	0	0	0
. abbreviation	0	1	0	0	0
__ in W	0	0	1	0	0

Table 17: Frequency Data from Punctuation Analysis of Pilot Subset Texts

Note: EOS means End Of Sentence; W means Word; dep means dependent or subordinate clause; S means Sentence.

Analysis: The underlying principle in punctuation analysis is the idea that punctuation reflects intonation, which is driven by syntactic structure (cf. Nunberg 1988) [81], Meyer (1987) [82]. Punctuation is therefore a reflection of syntactic structure, or an alternate means of

getting at syntactic structure. Punctuation is notoriously free in that rules for comma placement, for instance, are typically vague and underspecified. Because punctuation allows for options, it may also allow for individuality.

The hypothesis that syntactically-classified punctuation differentiates between writers can be tested statistically. Assuming the null hypothesis that there is no difference between the punctuation patterns of the pilot subset, what is the chance that these punctuation patterns come from the same author? Since the data is frequency of categories, the chi-square statistic is used, with the results shown in Table 18.

	01/09	01/16	01/80	09/16	09/80	16/80
$\chi 2$	32.664	74.409	31.212	90.049	24.852	52.165
p	0.0183	0.0001	0.0082	0.0001	0.052	0.0001

Table 18: Statistical Analysis of Punctuation Data in Pilot Subset Writers

The chances that the punctuation patterns from pairs of different writers are similar enough to conclude that the different writers are one and the same ranges from extremely small (1 in 10,000) to acceptably small (5 in 100). From these statistics, it can be inferred that punctuation patterns can differentiate between different writers.

The chi-square results in Table 19 relate to the null hypothesis that there is no difference between the punctuation patterns in QD2 and each of the writers in the Pilot Subset. Since the truth is that there is a difference between the author of QD2 and authors 001, 009 and 080, we expect a very low probability of no difference in these pairings, but a high probability of no difference in the pairing of QD2 and 016.

	01/qd2	09/qd2	80/qd2	16/qd2
$\chi 2$	58.846	88.674	48.003	18.904
p	0.0001	0.0001	0.0001	0.0909

Table 19: Statistical Analysis of Punctuation Data in QD and Pilot Subset

Table 19 shows that, as hoped for, there are very low probabilities of no difference when, in fact, the sources of the punctuation patterns really are different. When the sources of the punctuation patterns are the same, 016 and QD2, however, the probability of no difference fails the typical significance cut-off of $p < .05$. It would be nice if this p value were really high, but anything larger than .05 is acceptable in terms of the chi-square test. A similarity coefficient will have to be developed in order to deal specifically with issues of how similar two documents have to be in order to be classified as originating from one writer.

It is safe, however, to conclude that, at least in this forensically-similar task, frequency of syntactically-classified punctuation patterns is able to differentiate between different writers and cluster documents of one writer, in a statistically significant way.

Replication Results: The syntactically-classified punctuation technique has been replicated.

HYPOTHESIS 9: Abstract syntactic structures differentiate and identify authors.

Source: Chaski (1997a,1997b, 1998a) [60, 61, 63

Methodology: Parse text using a generalized phrase structure grammar.

Count structures and ratios between structures of related type.

Test for differences between texts statistically.

Tools: Knowledge of phrase structure grammars.
ALIAS® computer program.
Knowledge and use of _2 statistic.

ALIAS, Automated Linguistic Identification Authentication System, is an electronic parsing system which is designed to quantify the structures in a text. As a relational database, it consists of the components shown in Figure 1.

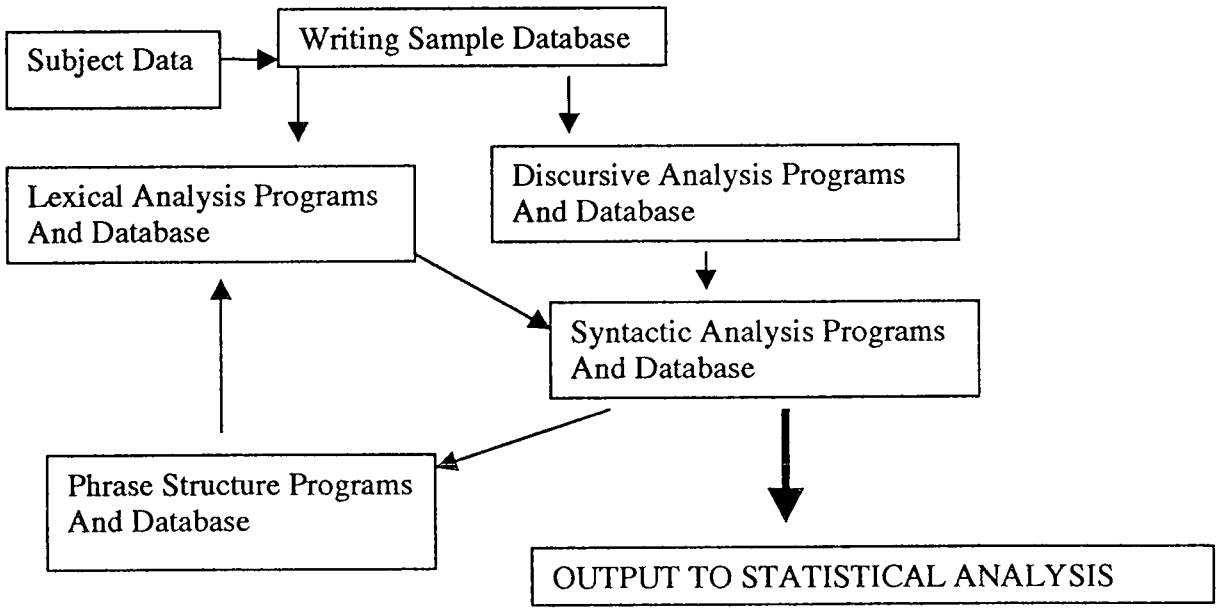

Figure 1: Components of ALIAS

These components perform the tasks and relate to each other as described in Figure 2. Each text passes from the Writing Sample Database through each component to statistical analysis.

Subject Info Database
stores sociological and dialectal information about each subject

Writing Sample Database
stores the texts written by each subject, keyed to Subject Information

Lexical Analysis Programs and Database
breaks text up into words,
assigns Part-Of-Speech (POS) labels
passes POS to Syntactic Analysis
sends quantification to statistical analysis

Discursive Analysis Programs and Database
breaks text up into sentences,

assigns discourse function,
passes sentences to Syntactic Analysis
sends quantification to statistical analysis

Syntactic Analysis Programs and Database
combines POS into bar and phrase levels,
combines Phrase Structures into sentences
sends quantification to statistical analysis

Phrase Structure Database
stores phrase structures,
parses to create phrases from POS
allows user to guide parsing decisions
sends quantification to statistical analysis

Output to Statistical Analysis

Figure 2: The Components and Functions of ALIAS

Statistical analysis enables us to determine identifying features, differentiating features, and idiolectal markers. A differentiating feature is a quantified, syntactic pattern which passes statistical testing of significant difference. An identifying feature is a quantified, syntactic pattern which fails statistical testing of significant difference. An idiolectal marker is a quantified, syntactic pattern which has both differentiating and identifying functions when submitted to significance testing.

Results in Tabular Format: Since ALIAS parses each word of a document, and each phrase of a document, may syntactic features are available for analysis. For brevity's sake, only data which illustrates the concepts of a differentiating feature, identifying feature, and idiolectal marker will be presented here.

Verb Phrase Features	Subject 016	Subject 080
mdl	17	18
v-prg	13	10
v-pas	8	1
v-pprt	16	1
mdl + v-pprt	1	0
v-pprt v-pas	1	0
v-neg inv	6	0
v-inf	27	11
v-inf pas	1	0
v v-ptl	71	44
vp[e]	1	1

Table 20: Raw Data of Frequencies of Verbal Features in Sets 016 and 080 of Pilot

Subject 016:	clauses/ sentence	phrases/ sentence
Text1	5.4	23.5
Text2	5.13	24.85
Text3	3.75	22.33

Table 21: Raw Data of Frequencies of Nodes Per Sentence in Set 016

	016-text1	016-text2	016-text3	080
pp[p np]	18	38	38	1
pp[variant]	1	7	3	5

Table 22: Raw Data of Frequencies of Prepositional Phrase Types in Sets 016 and 080

Analysis: The hypothesis that syntactic structures differentiate between writers can be tested statistically. Assuming the null hypothesis that there is no difference between the verb phrase patterns of the writers 016 and 080 from the pilot subset, as shown in

Table 20, what is the chance that these verb phrase patterns come from the same author? Since the data is frequency of categories, the chi-square statistic is used. When these frequencies are submitted to statistical testing, $\chi 2 = 19.739$, $p = .0318$. The probability of no difference (same origin) is very low, which in fact coincides with the fact that the documents were authored by different writers. Thus, verb phrase features function as a differentiating feature in this case.

On the other hand, the hypothesis that syntactic structures can identify or cluster documents written by the same writer can also be tested statistically. Assuming the null hypothesis that there is no difference between the complexity of sentences as measured by nodes per sentence in the writing of one author, as shown in Table 20, what is the chance that these nodes-per-sentence patterns come from the same author? Here we find a resounding failure of significant difference, $\chi 2 = .185$, $p = .9117$, which is just what we would expect. The probability of no difference is very high because, in fact, these documents do come from the same origin. Thus, sentential complexity in terms of nodes per sentence serves as an identifying feature in this case.

Finally, we need features which are able to distinguish between writers because they are used differently by different writers, but also identify documents because they are used consistently by each writer. The ratio of prepositional phrase types

pp[p np]
pp[p vp], pp[p p xp]

is a potential idiolectal marker which has both a differentiating and identifying function in the comparison of sets of documents. First, the notion of consistency across documents authored by one writer can be tested statistically. The data on prepositional phrases from Table 22 were run through the chi-square test to determine the chance of no significant difference between subject 016's prepositional phrase types.

	All 3 Texts of 016	Texts 1 and 2	Texts 2 and 3	Texts 3 and 1
$\chi 2$ =	2.225	1.294	1.417	.088
p =	.3288	.2553	.2339	.7667

Table 23: Statistical Analysis of Prepositional Types in Set 016

The probability of no difference between 016's texts 1,2,3 is very high, as expected, since these texts were authored by the same writer.

Second, the notion of idiolectal difference across writers can be tested statistically. The data on prepositional phrases from Table 22 with additional data from writer 080's texts were run through a chi-square test.

	016-1/080	016-2/080	016-3/080	016-all/080
$\chi 2$ =	4.13	1.84	5.464	7.04
p =	0.0421	0.175	0.0194	0.0706

Table 24: Statistical Analysis of Prepositional Types in Sets 016 and 080

The probability of no difference between 016's texts and 080's text is very low for two texts, as required, and relatively low for one text, since these texts were authored by different writers.

Replication: At this stage of research, more pilot subsets are being extracted from the Writing Sample Database in order to perform replications of the method on different writer sets. However, based on the results presented here we can conclude that syntactic analysis looks like a very promising approach.

4. 1 Summary of Empirical Testing Results

It is generally agreed among both forensic linguists and traditional document examiners that no conclusion can be based on a single attribute. The combination of attributes or results from many different techniques lead to the conclusion that a set of documents were authored by the writer of a particular known set or not authored by any of the suspects. In line with this principle, Table 25

shows how diasastrously dangerous many of the language-based author identification techniques are.

Hypothesis	Incorrectly Differentiates between	Incorrectly Clusters Together	Identifies SQD2 with
1: TTR	009/016 and 001/080	009 and 016; 001 and 080	001/080
2: V1	001/009/016 and 080	009 and 016	080
3: Readability Scores	-----	001 and 080, 009 and 016, 001, 009 and 016	001 009 016 or 080
4: Content Analysis	001/009 and 080/016/009 001/009/080/016 and 080	001 and 009 080, 016 and 009 all together	001/009 080
5: Spelling Errors	001/016 and 009/080	009 and 080	001 and 016 001 001/016
6: Grammar Errors	001/009/016 and 080	001/009/016	001/009/016
7: Sentence Complexity	009/016 and 001/009/080	001 009 and 080	001 009 016 or 080

1. These conclusions are not based on any quantification leading to probabilities.

Table 25: Errors from Common Author Identification Techniques

The danger of these techniques is that justice could be subverted because certain ideas about language use which are commonly held but empirically indefensible could lead to false identifications or false eliminations.

So the most important conclusion of my research, in my opinion, is the fact that techniques based on common misconceptions of

language use as a means of identifying authorship are unreliable, inaccurate and should not be admitted as scientific evidence. The underlying ideas about language use may be held by either the American high school graduate or the language expert, but they are not a reliable foundation for authorship identification in court

The empirical results of the Pilot Subset studies also demonstrated that not all language-based author identification techniques are misleading or dangerous. Two of these techniques -- punctuation patterns and syntactic structures-- yielded results which enable us to differentiate between authors while clustering documents from each author, as shown in Table 26.

Hypothesis	Correctly Differentiates between	Correctly Clusters Together
Syntactically-Classified Punctuation	001, 009, 016 and 080	3 texts of 16
Syntactic Analysis of Phrase Structure	001,009, 016 and 080	3 texts of 16

Table 26: Correct Results from Syntax-Based Techniques

While punctuation patterns may seem to be an obvious kind of textual phenomena which both the American high school graduate and the language expert would pay attention to, the way that punctuation was used in the empirical test requires knowledge of syntactic structures and statistics. So while any juror or judge may notice that one document contains lots of hyphens while another does not, any juror or judge may not notice that the hyphens in the one document are always syntactically conditioned in ways that are not available in the other document. In other words, even such an obvious feature as punctuation has to be handled in a non-obvious way in order to yield reliable results for author identification. Syntactic phrase structures, on the other hand, are the kind of phenomena which are not obvious to the American high school graduate or the language expert who has not been trained in syntactic theory and analysis.

To sum up, empirical studies of current language-based author identification techniques make two points clear:

1. techniques relying on common misconceptions about language are, predictably, unreliable;

2. techniques relying on linguistic science appear to accurately cluster and discriminate documents.

Legal conclusions can be drawn:

1. The jury can rely on its own common misconceptions about language to erroneously determine the authorship of documents without having an expert make their mistake more certain.

2. The jury may need an expert witness to help them **not** rely on common misconceptions about language.

3. The jury may need an expert as a rebuttal witness to help them discount the claims of other experts who rely on common misconceptions about language.

Scientific conclusions can be drawn:

1. The Daubert ruling is a great boon to all scientists who are seeking to develop forensic methods by applying the scientific techniques peculiar to their discipline.

2. The scientists' or language experts' integrity, when high, is absolutely key to the development of novel forensic applications basic science, and when low, is the sure road to junk science.

3. The limitations of real science, most often stated in statistical probability, are more honest than the grand conclusions of pseudo-science.

BIBLIOGRAPHY

1. Donaldson, Russell G. 1985. "Admissibility of evidence as to linguistics or typing style (forensic linguistics) as basis of identification of typist or author." *American Law Reports,* Annotated 36 ALR4th 598.

2. Menicucci, Jeffrey D. 1977. "Stylistics evidence in the trial of Patricia Hearst." *Arizona State Law Journal.*

3. Squires , Susan. 1997. "Linguist developing scientific method to identify authorship." *The Criminal Practice Report,* 11, 24: 460-464.

4. McMenamin, Gerald R. 1993. *Forensic stylistics.* Amsterdam: Elsevier.

5. *Black's Law Dictionary,* abridged Sixth edition. 1991. West Publishing Co.: St. Paul, MN.

6. Risinger, D.M., Denbeaux, M.P., and Saks, M.J. 1989. Exorcism of ignorance as a proxy for rational knowledge: the lessons of handwriting identification 'expertise.' *University of Pennsylvania Law Review,* 137: 731-787.

7. Risinger, D.M., and Saks, M.J. 1996. Science and nonscience in the courts: Daubert meets handwriting Identification expertise. *Iowa Law Review,* 82, 1: 21-74.

8. Hansen, Mark. 1997. Evidence Section. *ABA Journal,* May 1997:76-78.

9. Tiersma, Peter M. 1993. Linguistic issues in the law. *Language,* 69:1, 113-137.

10. Giannelli, Paul C. 1993a. Forensic science: Frye, Daubert and the Federal Rules. *Criminal Law Bulletin,* 26, 5, 428-436.

11. Imwinkelreid, Edward. 1997. Forensic science: Frye's general acceptance test vs. Daubert's Empirical Validation Standard-- "either...or" or "both...and"? *Criminal Law Bulletin,* 33,1: 72-84.

12. Johnson,Lynn R., Six, Stephen N., and Hamilton, Patrick A. 1997. Deciphering Daubert. *Trial,* November 1997: 71-78.

13. Huber, Peter W. 1991. *Galileo's revenge: Junk science in the courtroom.* New York: Basic Books.

14. Giannelli, Paul. 1993b. "Junk science": the criminal cases. *The Journal of Criminal Law and Criminology,* 84, 1, 105-128.

15. Hagen, Margaret A. 1997. *Whores of the court: The fraud of psychiatric testimony and the rape of American justice.* New York: Regan Books

16. Levi, Judith. 1994. Second edition. *Language and law: A bibibliographic guide to social science research in the U.S.A.* Teaching Resource Bulletin No.4. Chicago, Ill.: American Bar Association.

17. Crystal, David. 1995. Review of Forensic Stylistics. *Language,* 71, 2: 381-385.

18.Finegan, Edward. 1990. Variation in linguists' analyses of author identification. *American Speech,* 65, 4, 334-340.

19. Morton, A.Q. 1978. *Literary Detection.* London: Bowker.

20. Morton, A.Q. 1991a. Proper words in proper places. Department of Computing Science Research Report, R18, University of Glasgow.

21. Morton, A.Q. 1991b. The scientific testing of utterances. Cumulative sum analysis. *Journal of the Law Society of Scotland,* 357-359.

22. Morton, A. Q. and Michaelson, S. 1990. The Qsum plot. Report CSR-3-90 from the Department of Computer Science, University of Edinburgh, James Clerk Maxwell Building, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ.

23. Michaelson, S. , Morton, A. Q., and Hamilton-Smith, N. 1977. "To couple is the custom." Department of Computer Science, University of Edinburgh.

24. Michaelson, S. and Morton, A. Q. 1973. Positional stylometry. in A.J. Aitken, R.W. Bailey and (eds), *The computer and literary studies.* Edinburgh: Edinbugh University Press. pp. 69-83.

25. O'Brien, D.P. and Darnell, A.C. 1982. *Authorship puzzles in the history of economics: A statistical approach.* London: Macmillan Press Ltd.

26. Mosteller, Frederick and Wallace, David L. 1984. second edition. *Applied Bayesian and classical inference: The case of the Federalist Papers.* New York: Springer-Verlag.

27. Totty, R. N., Hardcastle, R.A., and Pearson, J. 1987. Forensic linguistics: the determination of authorship from habits of style. *Journal of the Forensic Science Society,* 27: 13-28.

28. Hardcastle, R.A. 1993. Forensic linguistics: an assessment of the CUSUM method for the determination of authorship. *Journal of the Forensic Science Society,* 33, 2: 95-106.

29. Sanford, Anthony J., Aked, Joy P., Moxey, Linda M., Mullin, James. 1994. A critical examination of assumptions underlying the cusum technique of forensic linguistics. *Forensic Linguistics,* 151-167.

30. Smith, M. W. A. 1989. Forensic stylometry: a theoretical basis for further developments of practical methods. *Journal of the Forensic Science Society,* 29, 1: 15-33.

31. Smith, Wilfred. 1994. Computers, statistics and disputed authorship. in Gibbons, John, ed, *Language and the law*. Longman: New York. pp. 374-413.

32. Holmes, David I. and Hilton, Michael L. 1993. Cumulative sum charts for authorship attribution: An appraisal. Forensic Linguistics Occasional Electronic Newsletter, Issue 2.

33. Hilton, M.L. and Holmes, D.I. 1993. An assessment of Cumulative Sum charts for authorship attribution. *Linguistic and Literary Computing*, 8, 2, 73-80.

34. Dahl, H. 1979. *Word frequencies of spoken American English.* Essex, CT: Verbatim.

35. Kucera, H. and Francis, W.N. 1967. *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

36. Foster, Donald W. 1989. *Elegy by W.S.: A study in attribution.* Newark: University of Delaware Press.

37. Bailey, Richard. 1969. Statistics and style: A historical survey. in Dolezel, Lubomir and Bailey, Richard W. (eds) *Statistics and style.* New York: American Elsevier Publishing Company, Inc., pp. 217-236.

38. Yule, G. Udny. 1938. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. *Biometrika* 30: 363-390.

39. Fucks, Wilhelm. 1952. On the mathematical analysis of style. *Biometrika* 39: 122-129.

40. Milic, Louis T. 1967. *A quantitative approach to the style of Jonathan Swift.* The Hague.

41. Yule, G. Udny. 1944. *The statistical study of literary vocabulary.* Cambridge: Cambridge University Press.

42. Herdan, G. 1955. A new derivation and interpretation of Yule's 'characteristic' K. *Journal of applied mathematics and physics* (ZAMP), VI: 332-334.

43. Herdan, G. 1966. *The advanced theory of language as choice and chance.* New York: Springer-Verlag.

44. Holmes, David I. 1994. Authorship attribution. *Computers and the Humanities* 28: 87-106.

45. Miller, George A. 1996. *The Science of Words.* Scientific American Library/HPHLP: New York.

46. Miron, Murray S. and Pasquale, Thomas A. 1978. Psycholinguistic analyses of coercive communication. *Journal of psycholinguistic research.* 7, 2: 95-120.

47. Miron, Murray S. 1990. Psycholinguistics in the courtroom. In Robert W. Rieber and William A. Stewart, eds., *The language scientist as expert in the legal setting: Issues in forensic linguistics.* New York: New York Academy of Sciences, 55-64.

48. Miron, Murray S. 1981. The resolution of disputed communication origins. In N.J. Lass, ed., *Speech and language: Advances in basic research and practice,* New York: Academic Press, 405-466.

49. Miron, Murray S. 1983. Content identification of communication origin. In R. Reiber, ed., *Advances in forensic psychology and psychiatry,* Norwood, NJ: Ablex. pp. 113-146.

50. Chomsky, Noam. 1957. *Syntactic structures.* The Hague.

51. Chomsky, Noam. 1965. *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

52. Stone, Philip J., Bales, Robert F., Namenwirth, J. Zvi, and Ogilvie, Daniel M. 1966. *The General Inquirer: A computer approach to content analysis.* Cambridge, MA: MIT Press.

53. Osgood, Charles E., May, William H., Miron, Murray S. 1975. *Cross-cultural universals of affective meaning.* Chicago,IL: University of Illinois Press.

54. Martindale, Colin and McKenzie, Dean. 1995. On the utility of content analysis in author attribution: The Federalist. *Computers and the Humanities* 29: 259-270.

55. Ellis, Barbara G. and Dick, Steven J. 1996. Who was 'Shadow'? The computer knows: Applying grammar-program statistics in content analyses to solve mysteries about authorship. *Journalism and Mass Communication Quarterly,* 73,4: 947-962.

56. Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

57. Strauss-Larsen, Jamie. 1993. Orality and literacy: A case study of how technology is changing the traditional models. M.A. Thesis. North Carolina State University.

58. Ellegard, A. 1962. *Who was Junius?* Stockholm: Almqvist & Wiksell.

59. Svartik, Jan. 1968. *The Evans statements: A case for forensic linguistics.* Stockholm: Almqvist & Wiksell

60. Chaski, Carole E. 1997a. Who wrote it? Steps toward a science of authorship. *National Institute of Justice Journal.* Washington, DC: US Department of Justice.

61. Chaski, Carole E. 1997b. An electronic parsing system for document authentication. International Association of Forensic Linguists Biannual Meeting. Durham, NC.

62. Chaski, Carole E. 1998a. Electronic parsing for idiolectal features in suspect documents. Linguistic Society of America Annual Meeting. New York, New York.

63. Chaski, Carole E. 1998a. An automated language-based authorship system for document authentication. Questioned Documents Section. American Academy of Forensic Sciences Annual Meeting. San Francisco, CA.

64. Pollard, Carl and Sag, Ivan A. 1987. *Information-Based Syntax and Semantics. Fundamentals, Volume 1.* Stanford: CSLI.

65. Pollard, Carl and Sag, Ivan A.1993. *Agreement, Binding and Control. Information-Based Syntax and Semantics, Volume 2.* Chicago: University of Chicago Press.

66. Douglas, John, Burgess, Ann W., Burgess, Allen G., and Ressler, Robert K. 1992. *Crime classification manual: A standard system for investigating and classifying violent crimes.* New York: Lexington Books.

67. Osborn, Albert S. 1910. first edition. *Questioned documents.* Rochester: Lawyer's Cooperative.

68. Osborn, Albert S. 1926. *The problem of proof.* Newark: Essex.

69. Osborn, Albert S. 1929. second edition. *Questioned documents.* Albany: Boyd.

70. Conway, James V. P. 1959. *Evidential documents.* Springfield,IL: Charles C. Thomas.

71. Hilton, Ordway. 1982. revised edition. *Scientific examination of questioned documents.* Boca Raton: CRC Press.

72. Harrison,Wilson R. 1958. *Suspect documents: Their scientific examination.* London: Sweet and Maxwell.

73. Eagleson, Robert. 1994. Forensic analysis of personal written texts: a case study. in Gibbons, John, ed, *Language and the law.* Longman: New York. pp. 362-373.

74. Pickett, Penelope O. 1993. Linguistics in the courtroom. *FBI Law Enforcement Bulletin,* October, 6-9.

75. Ciaffoni, Robert. 1994. Comparative stylistics and gramformprints: Their application in solving questioned writings. focusing on lawyer-logotyped wills. Manuscript.

76. Berry, Thomas Elliott. 1971. *The most common mistakes in English usage.* New York: MacGraw-Hill, Inc.

77. Goutsos, Dionysis. 1995. Review Article: Forensic Stylistics. *Forensic Linguistics* 2, 1: 99-113.

78. Baker, John Charles. Pace: A test of authorship based on the rate at which new words enter an author's text. *Literary and linguistic computing.* 3, 1: 36-39.

79. Ule, Louis. no date. The rare word fallacy. Manuscript.

80. Chaski, Carole E. 1996. Linguistic methods of determining authorship. National Institute of Justice Research Seminar.

81. Nunberg, Geoffrey. 1988. *The linguistics of punctuation.* Stanford: CSLI.

82. Meyer, Charles F. 1987. *A linguistic study of American punctuation.* New York: Peter Lang.

The Writing Sample Database was designed to take into account both general statistical sampling issues and linguistic performance. The decision factors for the writers (or experimental subjects) included the availability of subjects; writing as normal part of the subject's lifestyle; dialect similarity or dialect grouping; generally equivalent educational level; and representation of both genders and several ethnicities. Based on these factors, writing samples were collected from two groups: Criminal Justice majors at a community college; and Business and Nursing majors at a private 4-year college. Table 2 shows the sex, age and race distributions of subjects in the current Writing Sample Database.

The decision factors for the writing samples (or experimental tasks) included: genre or text-type parameters; similarity to actual types of questioned documents, e.g. suicide notes, threatening/anonymous letters, etc.; and emotional level and home dialect. We know that the social context and communicative goal of a message affect its form. There are differences between the speech and the writing of each individual, differences between language behavior at home and at work, differences between language in a letter to a friend and an essay [56, 57]. Based on these factors, subjects wrote, at their leisure, on ten topics, some of which are meant to elicit enough emotion to evoke the home dialect, while others are intended to elicit a more formal or workplace dialect. Topics are listed in Table 3.

SEX MALE48

AGE	Unreported	->19	20-25	26-30	31-40	41+	TOTALS BY RACE
Unreported Race	2	0	0	0	0	0	**2**
White	3	12	14	3	2	1	**35**
Black	0	3	0	1	0	1	**5**
Black Hispanic	0	0	0	1	0	0	**1**
Black Native Am	0	0	0	0	0	0	**0**
Hispanic	0	1	1	0	0	0	**2**
Native Am	0	3	0	0	0	0	**3**

TOTALS BY AGE	**5**	**19**	**15**	**5**	**2**	**2**	**48**

SEX FEMALE 44

AGE	Unreported	>19	20-25	26-30	31-40	41+	**TOTALS BY RACE**
Unreported Race	1	0	0	0	0	0	**1**
White	0	6	5	6	5	3	**25**
Black	0	6	5	1	3	1	**16**
Black Hispanic	0	0	0	0	0	0	**0**
Black Native Am	0	0	1	0	0	0	**1**
Hispanic	0	0	1	0	0	0	**1**
Native Am	0	0	0	0	0	0	**0**
TOTALS BY AGE	**1**	**12**	**12**	**7**	**8**	**4**	**44**

Table 2. Distributions of Subjects by Sex, Age and Race

Task ID	Topic
1.	Describe a traumatic or terrifying event in your life
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 3. Writing Topics for Writing Sample Database

APPENDIX 1: Writing Samples from the Pilot Subset

001-01

Giving birth to my 4th child, 3 mos too early I was in a detox center and premature labor began. First of all, I should state I was in a detox center so I could give birth to a healthy child. I was gripped with unbelievable terror at the thought that my child was coming that early I didn't feel like he would have the opportunity to survive because I was an alcoholic and a crack cocaine user through out the whole pregnancy. The hospital did what they could to save the child but because of his low birth weight and under developement he didn't stand a chance. The whole ordeal took 12 hours from the onset of labor until the actual time of death and he died in my arms. I was helpless and totally powerless to do anything to help or ease his suffering. The doctors said that he didn't suffer, but really how do they know!! At that moment in time I believe I would have given my own life to save his. But now as I think who would have taken care of him or my other small children. I'm a single parent of 3 children. I believe my son gave his life so I could live and that's how I go on and stay clean and chemical free.

001-02

Numerous people and events influence me everyday in different ways. As far as me returning to school, I guess it would have to be wanting a better quality of life for my children and myself. The only way that I knew how to accomplish this is to return to school; and continue my education and show my children how important an education is now, so they don't have to wait until they are adults to get their education. Also the current job market had a high impact on my decision to get a degree, because there are no jobs available that would allow me to support my family effectively. We needed some financial security that a job at McDonald can't provide.

001-03

My Career goals is to achieve a BA in Behavioral Science Although I don't view it that way. I take it systematicly one thing at a time and one step at a time. First I will recieve an AA in CJ May 96. Then I plan to switch to Wilmington College where I plan to earn my BA who knows may I go further and get a MA also. I hunger for the knowledge in this field because not only do I learn of the human condition and diversity of culture, I also learn of myself and how to handle every day problems. We are all connected by some mannerism either by our uniquness or likenesses, also there is a thin line between the two. I like knowing the why's and that there is not one answer to certain questions. The more I learn the more I realize I don't know so it keeps me coming back. I like systematic approaches and the deviations to problems and solutions. This field has broaden my awareness that allow for trial + error. Fairness and "that's just the way it is."

009-01

One of the most terrifying events in my life was being held at gunpoint and told to get in the car by two men. All I could feel was dying without Christ in my life. I had a chance to run or get in the car. I was scared. I knew if I dying I would to go to hell and had not made peace with God. I am from a Christian background.

So many things ran across my mind. All I could see was this big gun that looked as if it was a cannon. I got in the car, one drove and the other held the gun on me and told me not to look at them. The one guy told me if I looked he would kill me. By the way, the one that was doing all the talking didn't rape me, but made the other guy do it. I believe he was a pervert. I was too scared to cry but wanted the event to end. At that time, I lived in Baltimore and girls were being raped, killed and thrown out on the expressway or beltway. When he told me I should take you to New Jersey, I almost lost it. I remembered my background started praying. They finally let me go. He told me to get out and don't look back. I ran and ran until I reached an apartment with a light. No one would answer the door. I knocked on the door still no one would answer. I don't know how I arrived at my apartment, but I did. I jumped in the shower trying to wash his hands off but kept feeling his touch and remembering what had happened. I tried to tell my husband what happened but he was too high to listen. I didn't call the police because I felt I would be taken through the 3rd degree. I had seen it happen to too many women and nothing done. So I lived with it. I think about it sometimes now, but because Christ is in my life- that is what makes the difference! He has taken the hurt away.

009-02
I have been influenced by many people. A boss I had was very educated, independent, and aggressive. She was very successful and knew what she wanted and how to obtain it. She was a go-getter, not afraid to talk to anyone. When she appeared in a room, no matter what she was wearing, you could see the authority she had. Most women have to wear a suit to have that type of authority. My mother and father both have influenced me because they always succeeded at anything they went after. They taught me never to give up- "a winner is not a quitter" and a "quitter is not a winner." Anything you strive after you can obtain, if you work hard enough. Even though they were unable to receive a proper education, they instilled in me the importance of an education. Honesty and integrity as well as respecting other feelings were also important. There are other people who influenced me, especially those who have had great obstacles and other factors but still went on in spite of. There was a deaf lady that received a Master's Degree that influenced me because she had been a hearing person before which is much tougher than being born that way. She had developed a disease and lost her hearing but against all odds she received a Master's. According to her, she had no encouragement from outsiders but her family was very supportive. To me, this is most important. Family is an important factor in everyone life! Many more people would be successful if they only had family support.

009-03
My ultimate goal is receive a BS degree in Criminal Justice. With this degree my plans are to work extensively with juveniles and addicts. Since I have started I have mixed emotions about exactly what to do because I have found so many avenues to pursue in this field. I love people and concerned about their well being. Since I was involved in many things in the past but overcame them; I feel can be an asset to many people. Counseling has always been a desire but I had a

family and they were more important at the time. People have always felt comfortable talking to me and relating their problems. I feel comfortable talking to anyone. I never been afraid to start a conversation. Therefore, counseling would be ideal. Another career goal is to own a bookstore with coffee shop (Gourmet) and a boutique. I love to shop but I hate to see too many of the same kind. Boutiques are unique, since they usually only have one or two of the same item, so much different from a department store. I would like to return to my first goal, education is priceless. Many times jobs are not obtained due to lack of education. I always have told my siblings, "don't ever give a person an opportunity not to hire you because of the lack of education or qualification." My oldest daughter obeyed my advice and completed. My son enlisted in the army, married and then entered workforce. Now he is pursuing his career in criminal justice. My youngest daughter has enlisted in the army after in the workforce for a few years. Maybe she will also take my advice and pursue a career and attend College. More important than all of the above I must be a success in my ministry. I would like to be a success in leading many teens, or anyone hurting, to Christ. After all is done, career, family etc. we all must give an account to Jesus as to what have we done for him and with him when He was offered to them. Our goals are only temporal to get us through this life! Most important, where will you spend eternity. God Bless!

016-01

I guess my most terrifying feeling is not being here for my two sons. My own mother died when I was 30, and I've always thought that I've sheltered and protected my sons as much, or more, than mom did me, I was the youngest of 4, and if I left this world early, I'm not sure how my boys would function. Both emotionally and physically. Emotionally, we are a very close threesome, relying and depending almost solely on one another, with me being a focal point for problems they find themselves unable to deal with. We talk about everything together, and I always find it amazing when their peers say things like "my mom doesn't treat me like yours"- I treat my kids as people who need structoring, raising and guidance- not as kids who "belong" to me. I wonder -if I die- who my boys would hash over the week's happenings with. Who would they turn to for guidance and understanding- my family is of little help because I've raised my sons so differently- the boys father's family is of no help- they're far away and don't even know the two guys. Physically -my boys have been sheltered, once again- from the cruel realities of today's world. At the ages of 16 and 20, they are only now becoming financially responsible, I have raised them to respect a dollar, but they are only now beginning to learn where that dollar has to go before it can go where they want it to go. If I left my children now- they would be alone in that I have kept them mine- I have not involved them in financial matters, I have not forced them to accept and be with family members who do not see our "way"- my kids would survive -I have taught them that- but it would not be an easy survival- I worry for them- jobs are scarce, cost of living rises more each day- Being a parent is a very real fear.

016-02

I think my mother influenced me more than anyone. As a child, we were taught a lot of values but in ways that most kids couldn't pick up on. Like -we were seldom told "no"- we were told things like "if you choose to do such and such, these are the results, you make your decision. As teenagers we were given the choice to hang out where we wanted, with whom we wanted but we were told things like "if your grandmother sees you there, would she be proud and say "hi"?" Or, "you are who you're seen with"- We were also seldom threatened, she did just as she said she was going to do- we knew that if she said she was going to pour cold water on us next time we didn't get up out of bed on time- that is exactly what she would do- no second chance. Two stories that stick out in my mind are: she got tired of my sister and I arguing over who's turn it was to do the dishes; she said if we couldn't decide, she would solve the problem and decide for us- as kids, we never seem to learn, so the next nite, the same old s__t, and the next thing we knew- mom had opened up the window next to the table and thrown all of the dinner dishes out the window onto the lawn- she turned to us and said "now neither one of you have to do dishes- there are none left to wash- your only problem now is to explain this to your father when he gets home" (he was a truck driver.) The other thing I remember well is: I seldom "thought" to hang up my coat when I got home from school, it was always laying on a chair, or on the couch, -anywhere but where it should have been- She kept telling me to take care of it -finally she told me if I didn't, she was throwing it out in the snow. Well, one morning in January, I asked her where my coat was, and, you guessed it- in the Snowbank outside the kitchen door- left there from the nite before- I was born and raised in Houlton, Maine- in January, in Maine, it's pretty damn cold- Mom taught us to stand up for our beliefs, try to walk away from an arguement, and to treat others as you want to be treated. The other two people who have influenced my life are my 2 sons- I have raised them by myself and it has been interesting, heartbreaking, thankless, and one hell of an experience. But I wouldn't trade that experience for a ship full of hundred dollar bills. They have taught me to laugh from the inside, to look at the world from the ground up, and to never loose sight of who I am and who I'll be. Having those 2 has taught me to respect my own feelings, to show them (my feelings) in a way I can be comfortable with later- and to hold onto my goals- never loose sight of the future- the past is what made us what we are today- and mom was right "Someday I'll thank her for what she did".

080-01

The scariest thing in my life was when the doctor told me I had to have a hysterectomy because my pap smear revealed positive cancer cells. My fear and the unknowing were awful. Would I have to have chemotherapy or radiation? Would I lose my hair. Would I die, and if so, how much would I suffer? I guess he noticed the fear in my eyes and tried to assure me that the cells were probably localized, but I was not buying this. He tried to assure me and calm my fears by stating that by removing my uterus, the cancer cells would not spread. The two weeks waiting for the surgery were hell. How would my children be if I died? Who would be there for them? I loved them so much and wanted to see them grow into adults. Most of the time I was scared- couldn't concentrate and cried when I was alone. At other times, I felt guilty for being so selfish. I would scold

myself and tell myself that I had no control over this and it was out of my hands and I should just accept whatever happened. But the fear of the unknown is stronger than rational thought, and would rear its ugly head. Years later, I guess my scariest moment was unfounded- but who knows for sure? The scariest thing in my life so far has been the question of immortality.

080-02
My third grade teacher influenced me greatly. She was very intelligent, warm, and funny. She encouraged me and in so doing instilled confidence in me which up to that point was lacking. Because of her, I became a better student and proud of my accomplishments. Because of her quiet and praising manner, I loved going to school and tried harder to please her so she would bestow her warmth and praise on me. Through her guidance, I excelled that year, and became more aware of what I could achieve if I applied myself.

080-03
My career goal is to land a position where I could become free of working two jobs as I have in the past. I would like this to be a management position as I enjoy this. In addition, I am fond of travel, so this would be an asset as I am willing to relocate. Office management or human resource management are areas of interest to me. My goal is obtaining either of these positions with a corporation providing employee benefits. Primarily, however, I am interested in a Monday to Friday job that would provide an adequate salary so I could enjoy weekends.

SQD2
A lot of things anger me but nothing makes me really angry. I've pondered this question for a couple of hours and can't come up with one single factor. I can describe lots of small, irritating examples - but no one large "thing". Injustice makes me angry- treating all people the same in any system- people are all different- all circumstances are different- no one person is exactly like another- stereo typing people- that makes me angry- commercials on TV that ask for money to feed starving kids over seas makes me angry (Sally Struthers looks like she could give up a meal or two)- has anyone really looked in their own neighborhood lately? What about those kids down the street? Maybe they're hungry, too. People who are capable of working but dont - or won't- make me angry- kids who say "I can't" make me angry- people who live in perfect worlds- created by money- make me angry. Disease -espeically cancer- makes me angry. Cancer stole my mother at 52, and she never harmed a single living thing- and bore such pain, never complained- her death made me very angry- Families who don't appreciate one another make me angry. Wives who take advantage of their mate- and vice versa- make me angry. Our country's system of child support paying makes me angry- one person suffers, one person gains- and the kid gets nothing -is often the case. Or like my children- no support at all- and no help from welfare- because I lived with my parents, or because I "make too much money:- is that after taxes? No, that's before Uncle Sam takes his share- Incompetence in the work place makes me angry. If you can;t do the job- let someone who can do it, do it- Blacks who use "prejudice" like the term "thank-you" make me angry. Whites who can't envision a black president make me

angry. people who don't vote make me angry- Seaford's school system makes me angry. Kids who go to college and goof off, make me angry.