

STYLOMETRIC AUTHORSHIP ATTRIBUTION: A DETECTIVE ROMANCE

Patrick Juola, Ph.D.

Lauritis Professor of Teaching and Technology
Mathematics and Computer Science Department
Duquesne University, Pittsburgh, PA

1

About myself

- My degree reads “computer science”
- But like most DH practitioners, I’m pretty interdisciplinary
- I’m a forensic scientist, specifically a stylometrist
 - (You’re, a... *What?*)
- I also read a lot of detective novels...

2

What does this have to do with digital editions?

- Common to scholarship generally, whether computer science, forensic science, or scholarly publishing
- We have a mass of “stuff,” some of it possibly even correct, and our task is to make sense of it and present it in a clear and cogent manner.

3

3

Investigation and evidence

- Evidence is key to any investigation
 - Useful even (especially?) before trial
 - But only if it is what it purports to be
 - E.g., R. Austin Freeman wrote a detailed analysis of how to forge a fingerprint in 1905(!)
- First question for any evidence: *is it genuine?*

4

4

True for non-forensics as well

- If you’re writing a biography of Abraham Lincoln, you want to know what really happened.
- **This** might not be an authentic source.

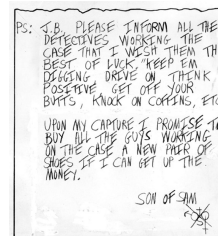


5

5

Questioned documents

- ... are a basic fact of life
- ... in fiction
- ... in scholarship
- ... and reality
- “Who wrote this?”



6

6

“Mysterious Note” – in fiction

- In Conan Doyle’s *The Hound of the Baskervilles*, Sir Henry Baskerville receives a note (made from words clipped from the newspaper) warning him “away from the moor.”
- But who wrote it?



7

7

Mysterious notes in scholarship

- The Epistle to the Hebrews
- The Donation of Constantine
- The Federalist Papers
- *Vortigern*, Shakespeare’s lost play
- The Hitler Diaries
- *Primary Colors*

8

8

Mysterious notes in forensics

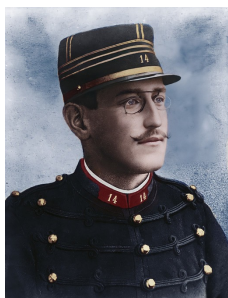
L’Affaire Dreyfus

Derek Bentley’s forged confession

The Unabomber Manifesto

The Jon Benet Ramsey ransom note

Ceglia v. Zuckerberg



9

9

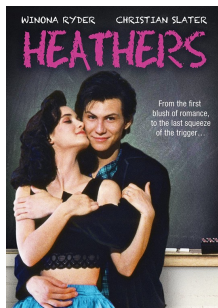
Questioned Documents

- Documentary evidence is important (increasingly so as we move online)
- Some crimes (forgery, document fraud, libel) can only be committed via documents
- Whole section of AAFS devoted to the physical examination of “questioned documents”

10

10

Is the suicide note real?



- **J.D.** : Um... ‘to me, though, suicide is the natural answer to the myriad of problems life has given me’.
- **Veronica Sawyer** : That’s good but Heather would never use the word myriad.
- **J.D.** : This is the last thing she’ll ever write; she’ll want to use as many 50-cent words as possible.
- **Veronica Sawyer** : She missed ‘myriad’ on the vocab test two weeks ago.
- **J.D.** : That only proves my point more. The word is a badge for her failures at school.

11

11

In real life (Chaski, 1995)

- You walk into a room, and there is your roommate’s dead body lying on the floor, with a suicide note on the computer, typed in MS-Word.
- Suicide or cunningly-concealed murder?

12

12

Dealing with these notes

- Traditionally, the realm of handwriting experts (today: FDE's)
- Can also involve typewriter experts
- However, modern documents (email, web pages, SMS) don't have "handwriting"
 - Any one 12pt Times Roman "A" looks identical to any other.
- What clues to authorship exist? How can we analyze such notes?

13

13

Basic approach (stylometry)

- "Language most shows a man; speak, that I may see thee." (Jonson, 1640)
- "Then said they unto him, say 'Shibboleth': and he said 'Sibboleth,' for he could not frame to pronounce it right. Then they took and slew him at the passages of Jordan." (Judges 12:6)

14

14

Sometimes it's obvious

- Back to the suicide note:
 - If it's in German, it wasn't me.
 - I know how to spell my own name.
 - I have my own verbal tics, not anyone else's.



15

15

Not so obvious?

- The dead roommate is named Angus Euan McScottishGuy, native to Glasgow.
- The note is dated 5/27/22 and reads, in part, "*I have been dishonored by my poor grades; the only way out is to take poison or throw myself under a semi.*"
- Anyone see any issues with this?

16

16

Analyzing the suicide note

- The note is dated 5/27/22 and reads, in part, "*I have been dishonored by my poor grades; the only way out is to take poison or throw myself under a semi.*"
- British writing:
 - Dates are written D/M
 - "dishonored" should have a 'u'
 - Students receive "marks", not "grades"
 - Goods are transported in "lorries," not trucks and certainly not "semis".
- Lots of reasons for suspicion here

17

17

Individual minds become individual writings

- Every person in this room thinks and acts slightly differently from every other one.
 - We have different interests,
 - ... different backgrounds,
 - ... different personalities,
 - ... different aptitudes
- All of which are reflected in language

18

18

About me:

- I was born in Seattle, WA, and grew up in Idaho/Oregon
 - I speak “Pacific Northwest”
 - “High muckamuck” – important person
 - “cougar” – mountain lion (*Puma concolor*)
 - ... but not perfectly
 - 4 years in mid-Atlantic
 - 7 in Colorado
 - 3 in UK (Oxford)
 - 20+ in Pittsburgh

19

19

Language as habits

The underlying linguistic theory is that all speaker/writers of a given language have their own personal form of that language, technically labeled an idiolect. **A speaker/writer's idiolect will manifest itself in distinctive and cumulatively unique rule-governed choices** for encoding meaning linguistically in the written and spoken communications they produce.

For example, in the case of vocabulary, **every speaker/writer has a very large learned and stored set of words** built up over many years. **Such sets may differ slightly or considerably from the word sets that all other speaker/writers have similarly built up**, in terms both of stored individual items in their passive vocabulary and, more importantly, **in terms of their preferences for selecting and then combining** these individual items in the production of texts.

20

20

Vocabulary habits

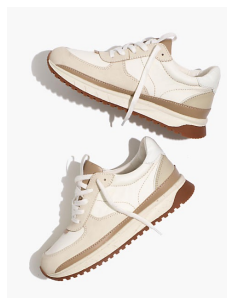


- What is this?
- Is this a couch?
 - Is this a sofa?
 - Is this a davenport?
 - Is this a chesterfield?
 - Is this a divan?
 - Is this a settee?

21

21

Vocabulary habits (2)



- What are these?
- Sneakers?
 - Trainers?
 - Running shoes?
 - Plimsolls?
 - Athletic shoes?

22

22

In real life (Shuy, 2001)

- Kidnapping case; primary evidence was ransom note
- “No Kops! Come alone!! [...] Put it in the green trash kan on the devil strip at the corner 18th and Carlson.”
 - Disregard misspellings; anyone can deliberately fake illiteracy
 - What on earth is a “devil strip”?

23

23

Devil strip

- The patch of grass between the sidewalk and the street
 - What do you call it?
 - Called “devil strip”
in Akron, OH



24

24

In real life (2) (Wellman, 1936)

- Found the word “touch” in several disputed documents
- Asked witness to write down dictated sentences, one included “touch”
- Witness spelled it as “touth”
- Offered (and accepted) as evidence that witness had written documents

25

25

Can work on phrases, too

- Any 7+ word phrase in English is likely to be unique
 - “My orange cat likes to eat fish”
 - o hits on Google as of 1 April 2021
 - “My orange cat likes to eat tuna”
 - “My orange cat loves eating tuna fish”
 - “My ginger cat loves eating tuna fish”
- Find a 7+ word phrase in common and you’ve proven common authorship

26

26

In real life (3)

- *Chevron Corp v. Donziger* (2014)
- Allegedly fraudulent decision; judge illegally substituted a decision ghost-written by plaintiffs in Ecuadorean court.
- Leonard/Juola analysis found nearly 40 lengthy phrases in decisions that were on the ghostwriter’s hard drive (and not in the record)

27

27

When it works, it’s great

- But it doesn’t always work. How often do you write about “touch,” regardless of you spell it?
 - Corpus studies estimate once/100,000 words.
- We need something more common and stable

28

28

Another type of cue

- The paradigmatic and systematic utilization of sesquipedalian lexical items can be an informative element of individual and idiosyncratic patterns of linguistic variation



Or, some people use big words

29

29

Word length as cue


- “Try to balance in your own mind the question whether the latter [text] does not deal in longer words than the former [text].... Some of these days spurious writings will be detected by this test.” (August de Morgan, 1851)
- Formally, calculate average word length and use first-semester stats (*t*-tests)

30

30

The issue


- Average word length is,... so-so.
- Like classifying people by height—almost everyone is average
- AWL isn't stable, either. I write differently to friends than to journals.
- If AWL won't work, what else will?



31

31

One solution



- Don't use averages
- Use multiple indicators separately
- Even if the suspect is average in one way, he's probably unusual in another

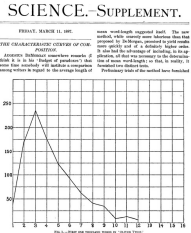
- A pair of eights is not the same as a six and a ten

32

32

Histograms of "features"


- Look at each length separately
- Can also be applied to other features than simple lengths
 - Parts of Speech
 - Common words
 - Punctuation
- Use stats to combine



33

33

An interesting feature set



Where is the dinner fork?

- "next to" the plate?
- "to the left of" the plate?
- "on the left of" the plate?
- "at the plate's left"?
- "left of" the plate?
- "right of" the salad fork?



- Or simply between the salad fork and plate?

34

34

Look at this slide


Three turtles sat on a log and a fish swam under them.


35

35

... and tell me what changed



Three turtles sat on a log and a fish swam under it.



36

36

Common words

- Strongly associated with “function” or “closed-class” words.
- Common (so lots of data)
- Little meaning (so lots of room for variation)
- Psycholinguistically “light,” so largely ignored *at a very fundamental level of processing*. (See Bransford, *et al.*, 1972) People focus on meaning, not form.

37

37

Literary: The Case of Oz XV



38

38

Historical synopsis

- Baum wrote *The Wonderful Wizard of Oz* in 1900; 13 more books by 1919.
- 15th book (“*Royal Book*”) unfinished at his death, Ruth Plumly Thompson took “fragmentary draft” and “unfinished notes” to complete and publish
- Published 1921 with Baum’s name on cover, “enlarged and edited” by Thompson

39

39

Historical synopsis

- Or was it?
- Alternate view: “*Royal Book*” was entirely Thompson’s work. Story about “unfinished notes” and Baum’s name created by publisher to strengthen sales.
- Thompson went on to publish nearly 20 more Oz books under her own name.

40

40

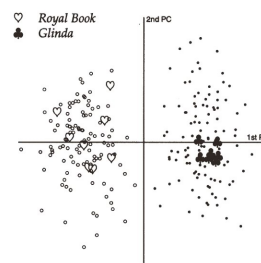
Binongo (2003) analysis

- Collect 50 “most frequent words”
- Candidates include *the, of, and, to, not, as, with, before, after, some, well, back, how*
 - *N.b.*, very “light” words
- Apply stats (PCA) to perform “dimensionality reduction,” essentially making 2D image from 50D data.

41

41

Baum is on right



So who wrote *Royal Book*?

42

42

In real life (Juola, 2012)

- I'm afraid I can't disclose much about this one.....
- Disputed will
- Analysis showed more similar vocabulary *and* more similar MWF vocabulary to claimant than to other testator writings
- My conclusion was that the will was forged

43

43

In real life (2) (Grant, 2013)

- 2009 fire; woman (AB) died
- Physical evidence was suspicious
- Text messages sent by AB seemed to prove she was alive that evening
- “Shift in texting style” at 12:07pm
 - “Xxx” and “xxx” became “X x”
- Were later messages sent by AB or her husband CB? Was AB already dead?

44

44

SMS vocabulary differs

Feature	Use by CB	Use by AB
“ad” for “had”	0	13
“dont” for “don’t”	0	9
“t” for “the”	1	15
“w” for “with”	3	10
“dnt” for “don’t”	8	0
“jst” for “just”	12	0
“wiv” for “with”	15	0
“4” for “for” (no space)	35	0

45

45

Grant’s analysis

- 18 features of known documents hand-identified
- Look for these features in questioned documents:
 - *Will do when i wake up, so tired. Need talk 2 chris. Still hav feelins **4him**. X*
- Grant concluded later messages inconsistent with AB’s style; CB later admitted setting fire and killing AB

46

46

“Case half solved”

- Method works by comparing “closeness” of documents
- More similar = “close” = more likely
- “Closed-class problem”: select the locked-room suspect
- “Open-class problem”: How about none-of-the-above?

47

47

Literature often open-class


- Who is Elena Ferrante?
 - Longlisted (2016) for Booker Prize
 - Known to be pseudonym, but for whom?
 - Is “Elena” even a woman?
- Typical English scholarship questions



48

48

Poe's early work?



- Analysis by Paul Collins (2013)
- Two prose works (1826)
- Author listed as Henry Poe (W. H. P.)
- Collins argues thematic and stylistic similarity to Edgar Allan's work
- Computational analysis (JGAAP) confirms: earliest known published prose by E.A.P.

49


Some well-known cases

- *The Federalist Papers*
- The “Unabomber Manifesto”
- Joe Klein’s *Primary Colors*
- Rowling’s *The Cuckoo’s Calling*
- *Ceglia v. Zuckerberg*

50

One of my cases (w/changes)

- Bilbo Baggins, a native of Mordor
- Anonymous on-line (newspaper) critic of Mordor government
- Applying for asylum in United States
- How to address his claim?



51

Evidence offered

අලිචායයන්ගේ
 විකල්පයක් ලෙසින්
 විද්වත්වීමට
 විකල්පයක් ලෙසින්

- ~10 signed articles (photocopies) from Mordor newspaper
- ~11 unsigned articles (web pages) from online paper
- If not him, who?

ඊළඟින් අලුත්වන
 රජයේ පත්ති මාලාව
 ඊළඟින් අලුත් වන

52

Key insight: ad-hoc KDs

- If it *could* be anyone, let it be anyone!
 - Bilbo’s writing will be most like his own!
- 5 named authors from political columns of online Elvish language newspaper
- 160 news articles, at least thirty articles per author

53

Two expected scenarios

- Yes
- No

Bilbo	Distractor #1
Distractor #1	Distractor #2
Distractor #2	Distractor #3
Distractor #3	Bilbo
Distractor #4	Distractor #4
Distractor #5	Distractor #5

- Bilbo closer than any competitor
- Rando off the street closer than Bilbo

54

Working hypothesis

- If Baggins were NOT the author of the questioned documents, there is only 1/6 chance he (KD) would be the closest author among the total author set to the QD
- Strengthen with multiple analyses
 - 1 analysis = 1/6
 - 2 analyses = 1/(6x6)
 - 5 analyses = 1/(6x6x6x6x6) = YAHTZEE!

55

55

The Cuckoo's Calling

- Published April 30, 2013
- Author named as Robert Galbraith
 - Acknowledged pseudonym
- Outed on Twitter as J.K. Rowling (author of *Harry Potter* series).
- Confirmation or disproof needed by newspaper



56

56

Rowling example

- Three distractor authors (contemporary female British crime novelists), total 4 candidates
- Elimination threshold used of top 2
- Four feature sets (word lengths, character 4-grams, word pairs, 100 most common words)
- Only Rowling not eliminated among candidates

57

57

Rowling example (2)

- How about none-of-the-above?
- If Rowling not author, chances of being in top 2 are 50% (2 of 4)
- Chances of being in top 2 4 times in a row are one in 16, about 6%. *Conclusion is "unlikely."*
- Rowling acknowledged authorship 13 July 2013



58

58

Software support

- JGAAP : Java Graphical Authorship Attribution Program
 - General-purpose analysis program
 - To be demonstrated
- Envelope : Proprietary J&A program to extend JGAAP to open-class verification
- BC: Proprietary J&A scalable phrasal overlap finder

59

59

JGAAP

- NSF-funded open-source freeware for authorship analysis experiments
- Offers many different choices for language experiments.
 - Let's play with it, shall we?

60

60



61