# Authorship Attribution Distance-based Methods

Jacques Savoy
University of Neuchâtel

Juola P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Love, H. (2002). *Attributing Authorship: An Introduction*, Cambridge University Press, Cambridge, 2002.

Craig H., Kinney A.F.(2009) Shakespeare, Computers, and the Mystery of Authorship, Cambridge, Cambridge University Press.

## Who is the author?

As possible authors, we have John F. Kennedy, Barack Obama, Abraham Lincoln. Attribute each text to its author.

**Text 1**: "Four score and seven years ago our fathers brought forth, upon this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal."

**Text 2**: "Yes, we can."

**Text 3**: "My fellow Americans, ask not what your country can do for you, ask what you can do for your country."

**Text 4**: "Ich bin ein Berliner"

## Authorship Attribution

- The Program…
  - **Problem & Context**
  - Examples
  - A single measurement
  - Multivariate analysis (restricted set of terms)
  - Distance-based approaches (*ad hoc*)

## Different Questions

Given a sample of texts known to be written by one of a set of authors,

- Question 1: *Closed-set.* Determine the author from a set of possible authors (e.g., political tract)
- Question 2: *Open-set.* Determine, if any, the author from the set of possible authors
- Question 3: *Verification*. Determine if the given author is really the correct one. Is it really Shakespeare? (Koppel *et al*., 2007)
- Question 4: *Profiling*. Determine pertinent attributes of the author (sex, age, education, psychological, …) (Pennebaker, 2011)

## The Output / The Data

- Only the most unlikely / probable author
  (or a ranked list)
- The style of the author (author's canon, stylistic traits)
- The assignment reliability
  - Minor vs. large impact: Attribution in the court room
    Forensic Linguistic, (Olsson, 2008)
    but also Pauline Epistles, The Book of Mormon, …
- Text sample
  - relatively large
  - balanced
  - high quality

## Beyond Simple AA

- Collaborative work (with?)
- Part of a play (e.g., a scene) (Craig & Kinney, 2009)
- Analysis character by character / dialogue
  Is Hamlet really a male character?
- Historical study of language change (diachronic linguistics) (Juola, 2003)
- Who is behind a politician?
- Profiling the author (Pennebaker, 2011), gender studies
- Plagiarism…
- Email (spam, fraud, propaganda) authentication

## How?

- Following St Jerome (347-420 AD)

1. if one book is inferior to the others
2. if the text contradicts the doctrine in author's other works
3. if the text is written in a different style, contains words and expressions not ordinarily found in the author's production
4. if passages quoting statements that were made or mentioning events that occurred after author's death

- *Comparative* basis

## Style

- Measurement of (aspects) of style
  "The stylometrist therefore looks for a unit of counting which translates accurately the 'style' of the text, where we may define 'style' as a set of measurable patterns which may be unique to an author"
  H. Holmes, Authorship Attribution, Computers and Humanities, 1994, p. 87
- Stylistic features (which ones?, how to select?, how many?)
  - Words, sequences of words, lemmas, *n*-grams, …
  - POS, sequence, proportions, …
  - Structural elements (e.g., layout, signature, logical structure, …)
- Hidden assumptions
  - The style is constant for an author in a given period
    and it differs from other authors

# Variations in Style

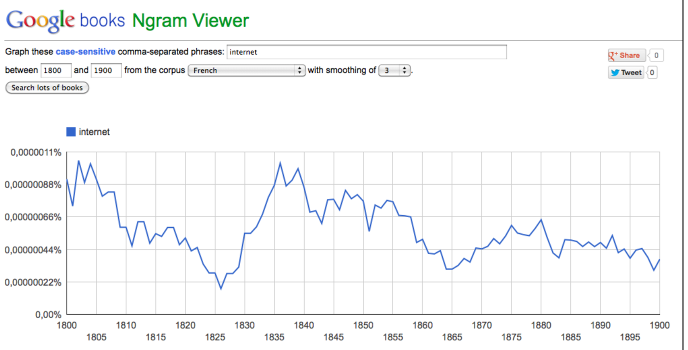
1. The village does not have a post office.
2. The village has no post office.
3. The village doesn't have a post office.
4. The village hasn't got a post office.
5. The village hasn't got no post office.
6. The village ain't got no post office.

Crystal, D. (2010). *A Little Book of Language.* Yale University Press

# Style

- Style is a function of
  - **Genre** (novel vs. poem, prose or verse)
  - **Author** (social, gender, age, education, native language, …)
  - **Period** (same time frame)
  - **Topic**
  - Type (spoken vs. written, web-based)
  - Audience (official vs. informal)
  - Editors / publishers
- Data quality (J. Rudman)
  - De-editing (page number, scene description, character's names)
  - Spelling normalization (one word = one spelling)

# Why Data Quality Matters



Google books Ngram Viewer

Graph these case-sensitive comma-separated phrases: internet

between 1800 and 1900 from the corpus French with smoothing of 3.

# Traditional Authorship Attribution

- Authorship attribution
  - External evidence (incipits, colophon, biographical evidence, earlier attributions, social world within which the work is created, …)
  - Internal evidence (self-reference, evidence from themes, ideas, beliefs, conceptions of genre, …) (St Jerome)
  - Bibliographical evidence
  - Historical, physical evidence (ink, handwritten, watermarks, multispectral imaging)

## Non-Traditional Authorship Attribution

- Stylometry (fingerprint)
  Computer Science & Statistics provide a quantitative tool
  - Single measure
  - Multivariate statistics
  - Distance-Based (similarity-based)
  - Machine Learning

  "when there are very many candidate authors, similarity-based methods are more appropriate than machine-learning methods."
  Koppel M., Schler J., Argamon S., Winter Y. (2012). The "Fundamentals Problem" of Authorship Attribution. *English Studies*, 93(3), 284-291.

## Authorship Attribution

- Overview
  - Problem & Context
  - **Examples**
  - A single measurement
  - Multivariate analysis (restricted set of terms)
  - Distance-based approaches (*ad hoc*)

## Notation

- Word type: distinct forms
- Word token: number of forms (« I saw a man with a saw »)
- Lemma: headword, entry in the dictionary
- V: vocabulaty used (word type)
- |V|: number of distinct word types
- $V_r$: vocabulary of terms appearing *r* times
- Hapax (hapax legomenon): word type appearing once
- $|V_1|$: number of hapax
- POS: Part-Of-Speech
- C: the corpus

## Notation

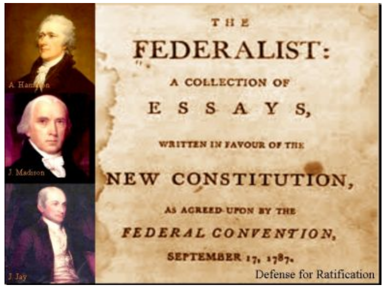
- $tf_{ij}$: absolute frequency of term i in text j (0, 1, … )
- $rtf_{ij}$: relative frequency of term i in text j ($0 \leq rtf_{ij} \leq 1$)
- $df_i$: document frequency (number of texts with term i)
- $a_j$: *j*th author, j = 1, 2, …, r
- m: number of selected features
- n: number of tokens in the corpus
- $n_j$: number of tokens of the *j*th text
  (or the size of the *j*th author profile)
- r: number of possible author

# Classical Examples

*The Federalist Papers*

Set of 85 essays written by *Publius*

In fact three possible authors:
A. Hamilton
J. Madison
J. Jay



Who wrote what? (Mosteller & Wallace, 1964)

# Classical Examples

- The *Federalist Papers* (Mosteller and Wallace, 1964)
  - A series of newspapers articles published in 1787-88 with the aim of promoting the ratification of the new US constitution. Papers written under the pseudonym “Publius”
  - Some are of known (and in some cases joint) authorship but others are disputed
  - Written by three authors, Jay (5), Hamilton (51) and Madison (14), three by Hamilton & Madison, 12 uncertain.
  - Pioneering stylometric methods were famously used by Mosteller and Wallace in the early 1960
  - It is now considered as settled
  - The *Federalist Papers* present a difficult but solvable test case, and are seen as a benchmark to test new ideas

# Federalist Papers: Zipf’s law

Vocabulary
7,860 word types (= |V|)
2,842 hapax (36%) (= $|V_1|$)
1,176 dis legomenon (= $|V_2|$)
both 51.1%

Size
167,190 tokens (= *n*)
123,669 Hamilton (74%)
43,521 Madison (26%)
10 most freq. -> 35.7%
50 most freq. -> 54.9%

| Rank | Word | Frequency |
|---|---|---|
| 1 | the | 14,200 |
| 2 | , | 10,333 |
| 3 | of | 9,467 |
| 4 | to | 5,751 |
| 5 | . | 4,065 |
| 6 | and | 3,849 |
| 7 | in | 3,591 |
| 8 | a | 3,247 |
| 9 | be | 3,025 |
| 10 | that | 2,221 |

# Federalist Papers: By Author

| | Hamilton | | Madison | |
|---|---|---|---|---|
| **Rank** | **Word** | **Freq.** | **Word** | **Freq.** |
| 1 | the | 10,293 | the | 3,907 |
| 2 | , | 7,508 | , | 2,825 |
| 3 | of | 7,149 | of | 2,318 |
| 4 | to | 4,498 | to | 1,253 |
| 5 | . | 2,998 | and | 1,168 |
| 6 | in | 2,782 | . | 1,067 |
| 7 | and | 2,681 | in | 809 |
| 8 | a | 2,476 | a | 771 |
| 9 | be | 2,270 | be | 755 |
| 10 | that | 1,679 | that | 542 |

## Hidden Questions: Tokenization

What is a word for you? And for the computer?

- Examples

  Richard Brown, 45-year old, is painting in New York
  I'll send you Paul's book
  John was prime minister to Henry VIII., permitting
  a final "take-it-or-leave-it" offer.
  Database system in the U.S.A.
  data base system in the US
  data-base system in the U.S.
  C|net, Micro$oft, and the IBM360, IBM-360, …

Sequence of letters and digits?

## Punctuation Marks?

The full stop (.) as a sentence length indicator.

"There is a strong personal element in the way people punctuate their writing. I know one novelist who puts commas in wherever possible. He writes sentences like this:

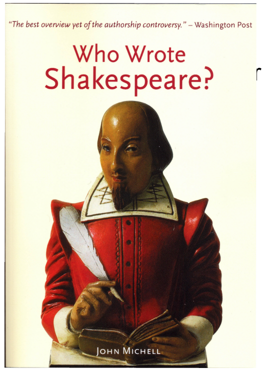
Fortunately, the bus was on time, so Sheema wasn't late for the concert.

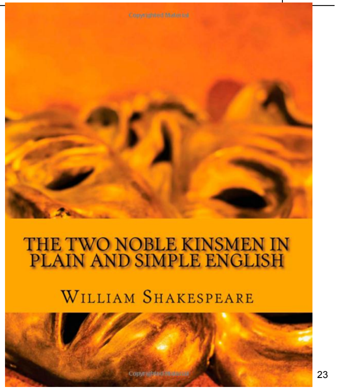
I know another who leaves them out whenever he can. He writes sentences like this:

Fortunately the bus was on time so Sheema wasn't late for the concert."

Crystal, D. (2010). *A Little Book of Language*. Yale University Press.

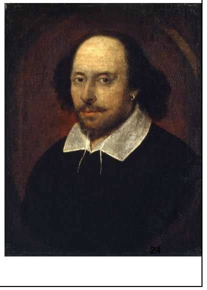
## Classical Examples


"The best overview yet of the authorship controversy." – Washington Post
Who Wrote Shakespeare?
John Michell



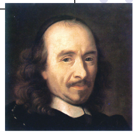
## Classical Examples

- Did Shakespeare write all of his plays?
  - Various authors including Bacon and Marlowe are said to have written parts or all of several plays
  - "Shakespeare" may even be a nom-de-plume for a group of writers?
- Plays written by more than one author
  - *Edward III* – Shakespeare? & Kyd?
  - *Two Noble Kinsmen* – Shakespeare & Fletcher
  - *Titus Andronicus* – Shakespeare & Peele?
  - *Henry VIII* – Shakespeare & Fletcher?
  - *Timon of Athens* - Shakespeare & Fletcher?



Craig H., Kinney A.F.: Shakespeare, Computers, and the Mystery of Authorship.
Cambridge University Press, 2009

## Classical Examples



- The debate *Molière* vs. *Corneille*?
  Jean Baptiste Poquelin (1622-1673)
  Pierre Corneille (1606-1684)
- *Psyché* (1671), both are authors
- Plays (comedies) from 1658
- Corneille needs money, well-known for his dramas (but cannot write comedies, and inferior genre)
- Pierre Louys (1919) (and Voltaire) indicates that Corneille was the real author based on the rhythmus, versification.

Labbé, D. (2009). Si deux et deux font quatre, Molière n' a pas écrit Dom Juan. Paris, Max Milo.

## Authorship Attribution

- Overview
  - Problem & Context
  - Examples
  - **A single measurement**
  - Multivariate analysis (restricted set of terms)
  - Distance-based approaches (*ad hoc*)

## Single Measurement

- Letter counts
- Word length
- Sentence length, too obvious and easy to manipulate
- Frequencies of letter pairs, strangely successful (*n*-gram)
- Distribution of words of a given length (in syllables), especially *relative frequencies*
- And what about the vocabulary growth and richness?
- Simple, but really effective?

## Vocabulary Richness

- Based on the idea that author's vocabulary is more or less constant
- Various measures
  - Type-token ratio
  - Simpson's index (the chance that two word arbitrarily chosen from text will be the same)
  - Yule's K (occurrence of a given word is a chance occurrence can be modelled as a Poisson distribution)
- But not stable for AA (Hoover, 2003), (Baayen, 2008)

$$Guiraud \quad R = \frac{|V|}{\sqrt{n}}$$

$$Sichel \quad S = \frac{|V_2|}{|V_1|}$$

$$Simpson \quad D = \frac{\sum_r r \cdot (r-1) \cdot V_r}{n \cdot (n-1)}$$

$$Yule \quad K = 10^4 \cdot \left( -\frac{1}{n} + \sum_r V_r \cdot \left(\frac{r}{n}\right)^2 \right)$$

## Letter counts

"What disturb me in Shakespeare's plays is the over-used of the letter "o". I can live with a lot of "e" or "I", but not a lot of "o". So, yes clearly, I prefer reading Marlowe."

## Letter Counts

- T. Merriam reports "of counting the letters in the 43 plays was the implausible discovery that the letter 'o' differentiates Marlowe and Shakespeare plays to an extent well in excess of chance" (used also the letter 'a')
- Frequency less than 0.0078, 6 plays of Marlowe Frequency greater than 0.0078, 36 plays of Shakespeare

T. Merriam: Letter Frequency as a Discriminator of Authors. *Notes & Queries*, 239, 1994, p. 467-469.
T. Merriam: Heterogeneous Authorship in Early Shakespeare and the Problem of *Henry V*. *Literary and Linguistic Computing*, 13, 1998, p. 15-28.

## Authorship Attribution

- Overview
  - Problem & Context
  - Examples
  - A single measurement
  - **Multivariate analysis** (restricted set of terms)
  - Distance-based approaches (*ad hoc*)

## Multivariate Analysis

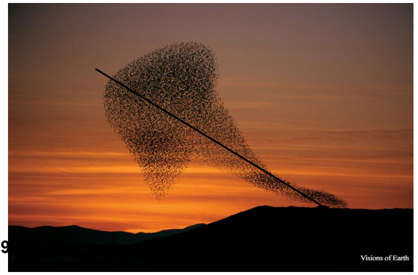
- Thanks to computers it is now possible to collect large numbers of different measurements, of a variety of features
- Variants of multivariate analysis
  - Principal components analysis (PCA)
  - Correspondence analysis (CA)
  - Cluster analysis
- Tools to **visualize** the data (better than reading a lexical table)
- Variables = features = word types or lemmas
- Objects = text excerpts

# Representation

PCA: explains the data using fewer variables

Explaining the max. of the **variability**

A cloud of birds in 3D → 2D (→ 1D)



See (**Binongo & Smith, 1999**
(Craig & Kinney, 2009)

# PCA: Input

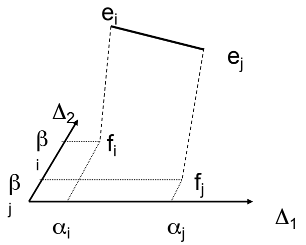
- Small lexical table with 4 texts (authors) and 5 words

| | A | B | C | D |
|---|---|---|---|---|
| the | **15** | **30** | 5 | 12 |
| a | **9** | **20** | 3 | 8 |
| I | 2 | 4 | **10** | 4 |
| my | 2 | 2 | **13** | 4 |
| of | 2 | 6 | 4 | 3 |

- B is “twice” A
- A and B have more determiners “the” and “a” than other words
- C used more “I” and “my”
- D is the style of the average
- Visualize this data (apply PCA (normalize))

# Principal Component Analysis

- PCA generate a smaller ordered set of new variables (principal components) uncorrelated (latent factors)
- “principal components” are computed by calculating the *correlations* between all the terms, then grouping them into sets that show the most correspondence



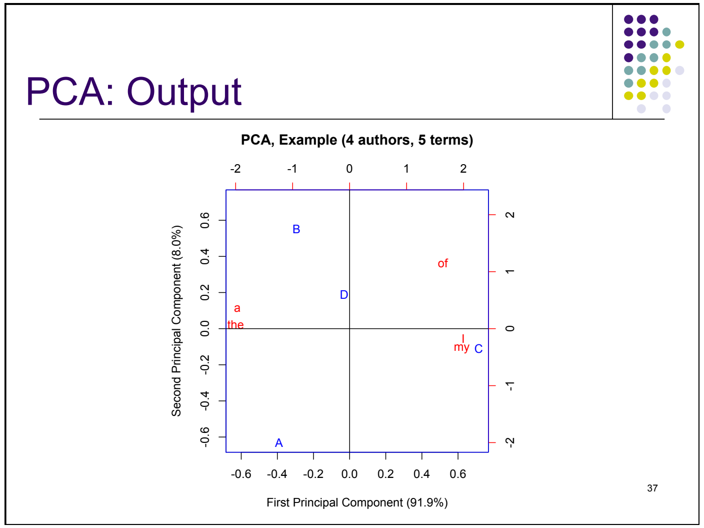
We will define a projection plane (defined by the lines $\Delta_1$ and $\Delta_2$, *perpendicular* (no correlation)) to represent the objects ($e_i$, $e_j$) and conserving the real distance d($e_i$, $e_j$).
Focus: dispersion

# PCA: Input

- Small lexical table with 4 authors (texts) and 5 words

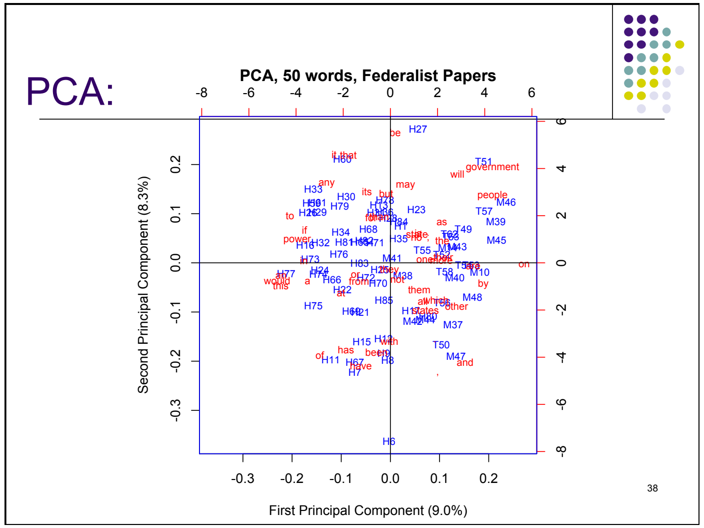
| | A | B | C | D |
|---|---|---|---|---|
| the | **15** | **30** | 5 | 12 |
| a | **9** | **20** | 3 | 8 |
| I | 2 | 4 | **10** | 4 |
| my | 2 | 2 | **13** | 4 |
| of | 2 | 6 | 4 | 3 |

- B is “twice” A
- A and B have more determiners “the” and “a” than other words
- C used more “I” and “my”
- D is the style of the average



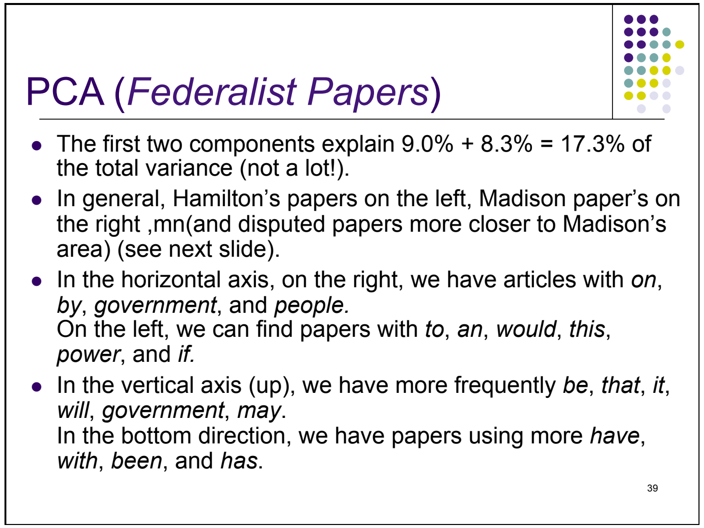
## PCA: Output

37



## PCA:

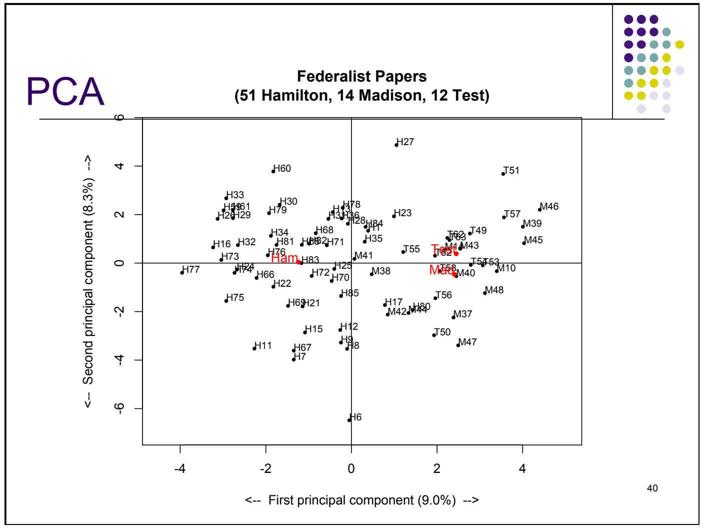
38



## PCA (*Federalist Papers*)

- The first two components explain 9.0% + 8.3% = 17.3% of the total variance (not a lot!).
- In general, Hamilton's papers on the left, Madison paper's on the right ,mn(and disputed papers more closer to Madison's area) (see next slide).
- In the horizontal axis, on the right, we have articles with *on*, *by*, *government*, and *people*.
  On the left, we can find papers with *to*, *an*, *would*, *this*, *power*, and *if.*
- In the vertical axis (up), we have more frequently *be*, *that*, *it*, *will*, *government*, *may*.
  In the bottom direction, we have papers using more *have*, *with*, *been*, and *has*.

39



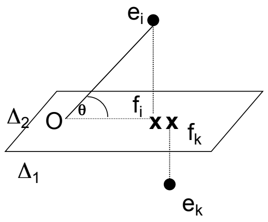
## PCA

40

## Principal Component Analysis

Visual and real distance.

Having two points $f_i$ and $f_k$ close together in the PC1 and PC2 plan does not mean that the corresponding $e_i$ and $e_k$ points are also close together.



PCA could be useful in your context,
- to visualize
- to synthesis your data!

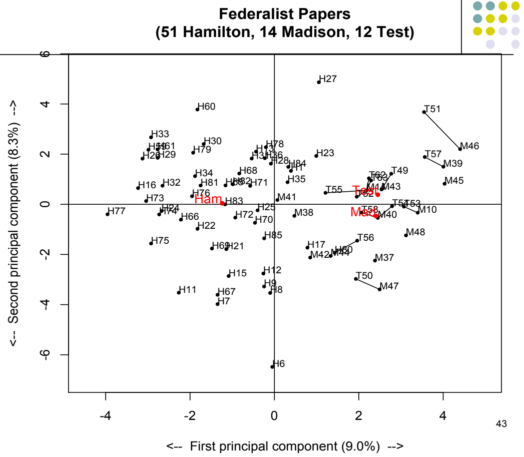
## Nearest Neighbour

- But we can imagine a simple attribution method… Find the text / author profile having the smallest distance with the representation of a disputed text,
- Testing instance Q:
  - Compute similarity between Q and all other texts / author profiles
  - Assign Q the category of the most similar example (1-NN)
- Simple to apply. The system does not really learn the different styles.
- Nearest neighbor method depends on a distance measure

## PCA

11 articles assigned to Madison (good)

Paper #56 assigned to Hamilton



## Authorship Attribution

- Overview
  - Problem & Context
  - Examples
  - A single measurement
  - Multivariate analysis (restricted set of terms)
  - **Distance-based approaches**
    Delta
    Chi-square
    Kullback-Leibler
    Vocabulary
    Labbé's distance

# Burrows' Delta

- Based on on the *m* most (*m* = 50, …) frequent words
  (+ POS for some types such as *to*, *in*)
  "frequency-hierarchy for the most common words in a large group of suitable texts" (p. 269)
- Compute a Z-score value for each word
  - for each word type $w_i$ , i = 1, …, in a text $D_j$, compute the relative frequency $rtf_{ij}$ (in ‰)
  - $\mu_i$ mean in the reference corpus
  - $\sigma_i$ standard deviation

$$Z\ score(w_{ij}) = \frac{rtf_{ij} - \mu_i}{\sigma_i}$$

Burrows, J. F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267-287.

# Top 50 Most Frequent Words
(Hamilton & Madison)

the
,
of
to
.
and
in
a
be
that

it
is
which
as
by
;
this
would
have
or

for
not
will
with
from
their
an
are
they
states

been
on
government
may
state
all
power
its
but
has

other
if
at
any
than
more
no
there
them
one

# Burrows' Delta

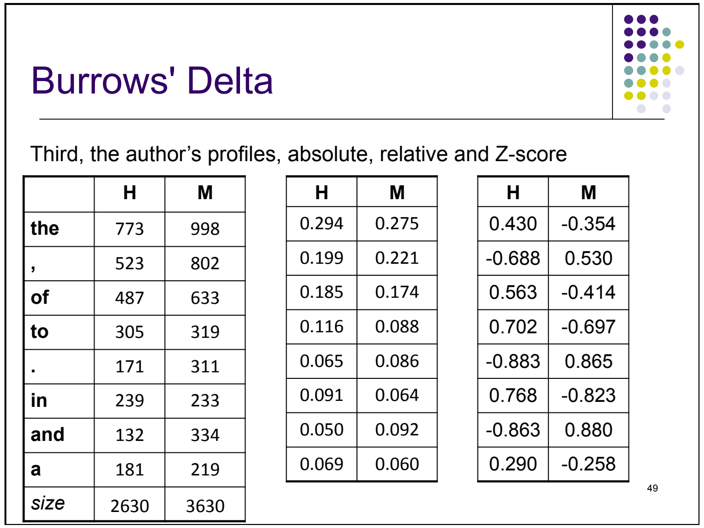
First compute the absolute frequencies in our example.

| | H59 | H60 | H61 | H62 | M37 | M38 | M47 | M48 |
|---|---|---|---|---|---|---|---|---|
| **the** | 177 | 224 | 152 | 220 | 230 | 273 | 328 | 167 |
| **,** | 133 | 152 | 104 | 134 | 192 | 234 | 219 | 157 |
| **of** | 112 | 145 | 100 | 130 | 159 | 189 | 187 | 98 |
| **to** | 73 | 87 | 61 | 84 | 84 | 117 | 64 | 54 |
| **.** | 45 | 47 | 32 | 47 | 75 | 95 | 85 | 56 |
| **in** | 62 | 79 | 47 | 51 | 63 | 62 | 62 | 46 |
| **and** | 34 | 36 | 25 | 37 | 101 | 95 | 87 | 51 |
| **a** | 49 | 53 | 35 | 44 | 57 | 92 | 35 | 35 |
| *size* | *636* | *770* | *521* | *703* | *904* | *1065* | *1032* | *629* |

# Burrows' Delta

Then the relative frequencies, and the mean and stdev

| | H59 | H60 | H61 | H62 | M37 | M38 | M47 | M48 | μ | σ |
|---|---|---|---|---|---|---|---|---|---|---|
| **the** | 0.278 | 0.291 | 0.292 | 0.313 | 0.254 | 0.256 | 0.318 | 0.266 | *0.283* | *0.024* |
| **,** | 0.209 | 0.197 | 0.200 | 0.191 | 0.212 | 0.220 | 0.212 | 0.250 | *0.211* | *0.018* |
| **of** | 0.176 | 0.188 | 0.192 | 0.185 | 0.176 | 0.177 | 0.181 | 0.156 | *0.179* | *0.011* |
| **to** | 0.115 | 0.113 | 0.117 | 0.119 | 0.093 | 0.110 | 0.062 | 0.086 | *0.102* | *0.020* |
| **.** | 0.071 | 0.061 | 0.061 | 0.067 | 0.083 | 0.089 | 0.082 | 0.089 | *0.075* | *0.012* |
| **in** | 0.097 | 0.103 | 0.090 | 0.073 | 0.070 | 0.058 | 0.060 | 0.073 | *0.078* | *0.017* |
| **and** | 0.053 | 0.047 | 0.048 | 0.053 | 0.112 | 0.089 | 0.084 | 0.081 | *0.071* | *0.024* |
| **a** | 0.077 | 0.069 | 0.067 | 0.063 | 0.063 | 0.086 | 0.034 | 0.056 | *0.064* | *0.015* |



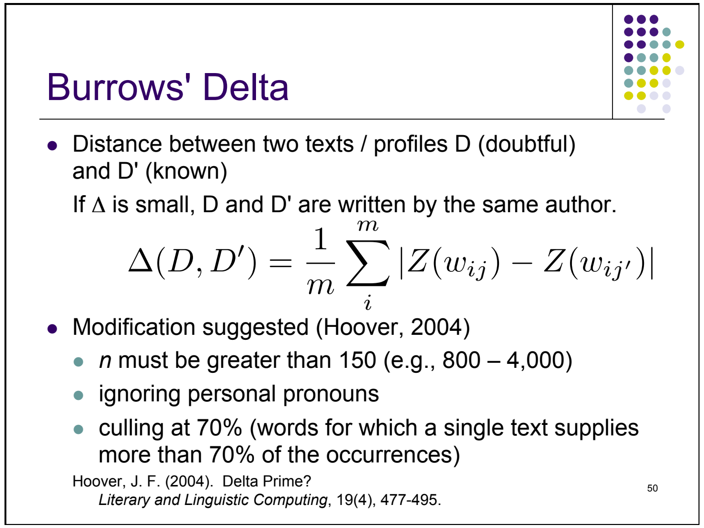
## Burrows' Delta

Third, the author's profiles, absolute, relative and Z-score

| | H | M |
|---|---|---|
| **the** | 773 | 998 |
| **,** | 523 | 802 |
| **of** | 487 | 633 |
| **to** | 305 | 319 |
| **.** | 171 | 311 |
| **in** | 239 | 233 |
| **and** | 132 | 334 |
| **a** | 181 | 219 |
| *size* | 2630 | 3630 |

| H | M |
|---|---|
| 0.294 | 0.275 |
| 0.199 | 0.221 |
| 0.185 | 0.174 |
| 0.116 | 0.088 |
| 0.065 | 0.086 |
| 0.091 | 0.064 |
| 0.050 | 0.092 |
| 0.069 | 0.060 |

| H | M |
|---|---|
| 0.430 | -0.354 |
| -0.688 | 0.530 |
| 0.563 | -0.414 |
| 0.702 | -0.697 |
| -0.883 | 0.865 |
| 0.768 | -0.823 |
| -0.863 | 0.880 |
| 0.290 | -0.258 |



## Burrows' Delta

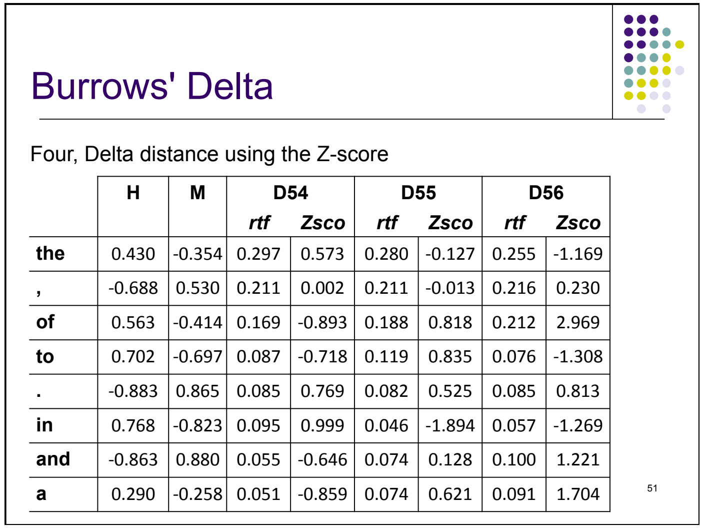
- Distance between two texts / profiles D (doubtful) and D' (known)

  If Δ is small, D and D' are written by the same author.

$$\Delta(D, D') = \frac{1}{m} \sum_{i}^{m} |Z(w_{ij}) - Z(w_{ij'})|$$

- Modification suggested (Hoover, 2004)
  - *n* must be greater than 150 (e.g., 800 – 4,000)
  - ignoring personal pronouns
  - culling at 70% (words for which a single text supplies more than 70% of the occurrences)

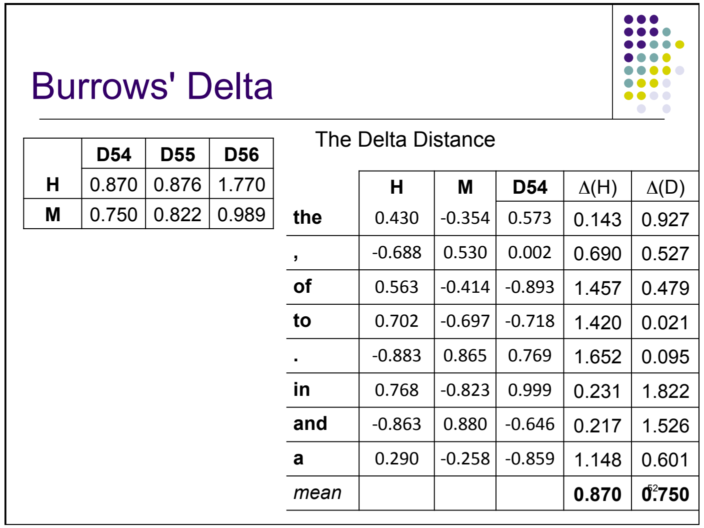
Hoover, J. F. (2004). Delta Prime?
*Literary and Linguistic Computing*, 19(4), 477-495.



## Burrows' Delta

Four, Delta distance using the Z-score

| | H | M | D54 | | D55 | | D56 | |
|---|---|---|---|---|---|---|---|---|
| | | | *rtf* | *Zsco* | *rtf* | *Zsco* | *rtf* | *Zsco* |
| **the** | 0.430 | -0.354 | 0.297 | 0.573 | 0.280 | -0.127 | 0.255 | -1.169 |
| **,** | -0.688 | 0.530 | 0.211 | 0.002 | 0.211 | -0.013 | 0.216 | 0.230 |
| **of** | 0.563 | -0.414 | 0.169 | -0.893 | 0.188 | 0.818 | 0.212 | 2.969 |
| **to** | 0.702 | -0.697 | 0.087 | -0.718 | 0.119 | 0.835 | 0.076 | -1.308 |
| **.** | -0.883 | 0.865 | 0.085 | 0.769 | 0.082 | 0.525 | 0.085 | 0.813 |
| **in** | 0.768 | -0.823 | 0.095 | 0.999 | 0.046 | -1.894 | 0.057 | -1.269 |
| **and** | -0.863 | 0.880 | 0.055 | -0.646 | 0.074 | 0.128 | 0.100 | 1.221 |
| **a** | 0.290 | -0.258 | 0.051 | -0.859 | 0.074 | 0.621 | 0.091 | 1.704 |



## Burrows' Delta

| | D54 | D55 | D56 |
|---|---|---|---|
| **H** | 0.870 | 0.876 | 1.770 |
| **M** | 0.750 | 0.822 | 0.989 |

The Delta Distance

| | H | M | D54 | Δ(H) | Δ(D) |
|---|---|---|---|---|---|
| **the** | 0.430 | -0.354 | 0.573 | 0.143 | 0.927 |
| **,** | -0.688 | 0.530 | 0.002 | 0.690 | 0.527 |
| **of** | 0.563 | -0.414 | -0.893 | 1.457 | 0.479 |
| **to** | 0.702 | -0.697 | -0.718 | 1.420 | 0.021 |
| **.** | -0.883 | 0.865 | 0.769 | 1.652 | 0.095 |
| **in** | 0.768 | -0.823 | 0.999 | 0.231 | 1.822 |
| **and** | -0.863 | 0.880 | -0.646 | 0.217 | 1.526 |
| **a** | 0.290 | -0.258 | -0.859 | 1.148 | 0.601 |
| *mean* | | | | **0.870** | **0.750** |

## Selection in Delta

Selection the *k* most frequent forms, we can find

1. Referential: articles and pronouns
2. Temporal / modal: auxiliary verbs and some adverbs
3. Connective: conjunctions, prepositions, relative pronouns
4. Modificatory: adjectives , adverbs

## Chi-Square Approach

- How many terms do we need to take account?

- A set of terms (very frequent) defined *a priori*?
- Define *m* number of features,
  as a *k*-limit, meaning that the selected terms must appear in at least *k* documents writing by each author (*df*-based)
- Low *k* value means a larger number of terms
- Large *k* value implies a smaller set of features
  (limit: the total number of articles writing by a single author)
- Guarantee that each cell is not empty (smoothing is not needed)

## Chi-Square Approach

- Distance between Q query text and $A_j$ the author profile of author j

$$\chi(Q, A_j) = \sum_{i=1}^{m} \frac{(rtf_{iq} - rtf_{ij})^2}{rtf_{ij}}$$

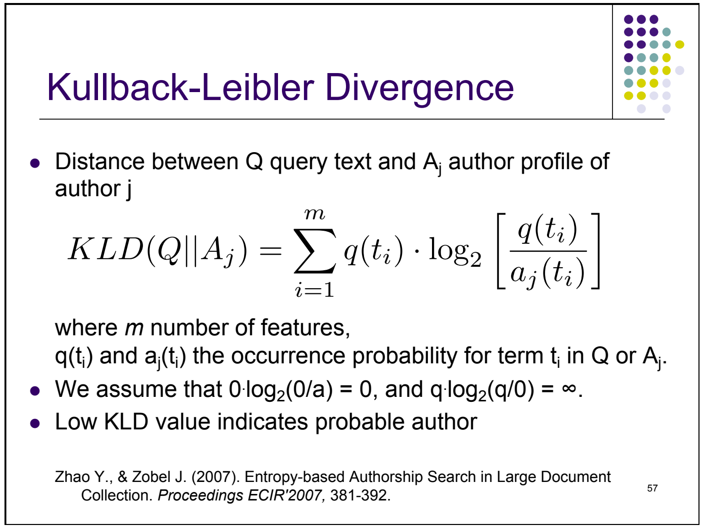
  where *m* number of features,
  $rtf_{iq}$ and $rtf_{ij}$ the occurrence probability for term $t_i$ in Q or $A_j$.
- Low $\chi$ value indicates probable author.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary & Linguistic Computing*, 22(3), 251-270.

## Kullback-Leibler Divergence

- We can define *a priori* a set of very frequent words appearing in a given language
- Stopword list in information retrieval (search engines)
  Zhao & Zobel: 369 terms (e.g., *the*, *we*, *is*, *not*, *became*, ….)
  Italian language: 399 terms (e.g., *del*, *essi*, *non*, *volta*,…)
- For each word, we can estimate the occurrence probability $q(t_i)$ and $a_j(t_i)$ for term $t_i$ in Q or $A_j$
- Compute the distance between the distribution in the query text Q and the distribution obtained from each author profile

Zhao Y., & Zobel J. (2007). Entropy-based Authorship Search in Large Document Collection. *Proceedings ECIR'2007*, 381-392.



## Kullback-Leibler Divergence

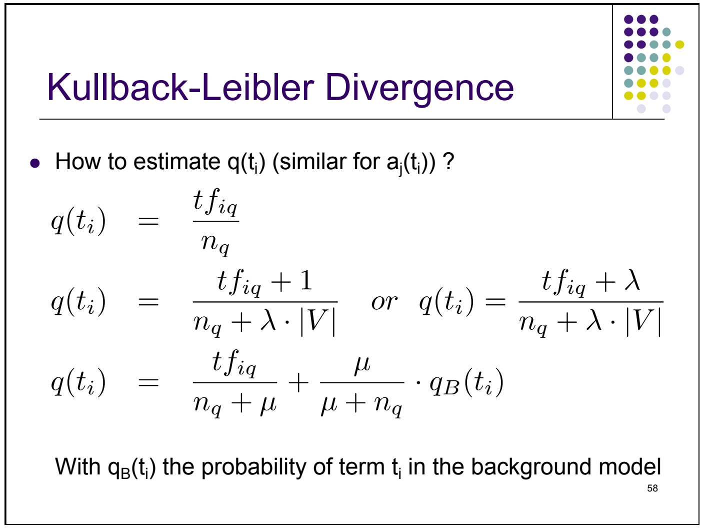
- Distance between Q query text and $A_j$ author profile of author j

$$KLD(Q||A_j) = \sum_{i=1}^{m} q(t_i) \cdot \log_2 \left[ \frac{q(t_i)}{a_j(t_i)} \right]$$

where *m* number of features,
$q(t_i)$ and $a_j(t_i)$ the occurrence probability for term $t_i$ in Q or $A_j$.

- We assume that $0 \cdot \log_2(0/a) = 0$, and $q \cdot \log_2(q/0) = \infty$.
- Low KLD value indicates probable author

Zhao Y., & Zobel J. (2007). Entropy-based Authorship Search in Large Document Collection. *Proceedings ECIR'2007*, 381-392.



## Kullback-Leibler Divergence

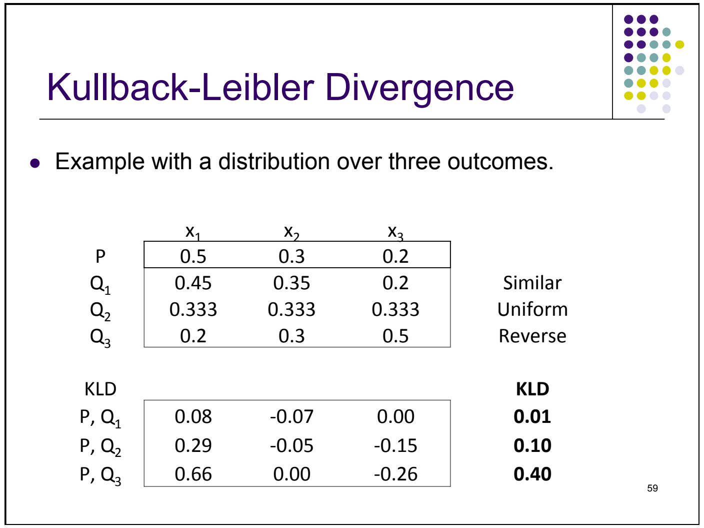
- How to estimate $q(t_i)$ (similar for $a_j(t_i)$) ?

$$q(t_i) = \frac{tf_{iq}}{n_q}$$

$$q(t_i) = \frac{tf_{iq} + 1}{n_q + \lambda \cdot |V|} \quad or \quad q(t_i) = \frac{tf_{iq} + \lambda}{n_q + \lambda \cdot |V|}$$

$$q(t_i) = \frac{tf_{iq}}{n_q + \mu} + \frac{\mu}{\mu + n_q} \cdot q_B(t_i)$$

With $q_B(t_i)$ the probability of term $t_i$ in the background model

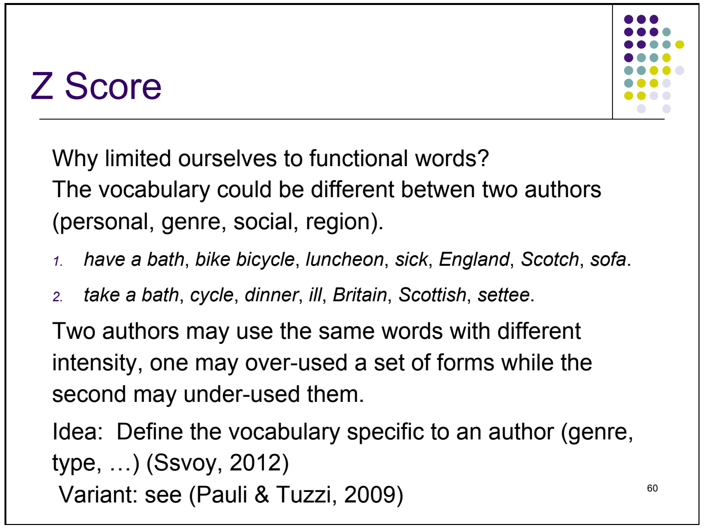


## Kullback-Leibler Divergence

- Example with a distribution over three outcomes.

| | $x_1$ | $x_2$ | $x_3$ | |
|---|---|---|---|---|
| P | 0.5 | 0.3 | 0.2 | |
| $Q_1$ | 0.45 | 0.35 | 0.2 | Similar |
| $Q_2$ | 0.333 | 0.333 | 0.333 | Uniform |
| $Q_3$ | 0.2 | 0.3 | 0.5 | Reverse |

| KLD | | | | **KLD** |
|---|---|---|---|---|
| P, $Q_1$ | 0.08 | -0.07 | 0.00 | **0.01** |
| P, $Q_2$ | 0.29 | -0.05 | -0.15 | **0.10** |
| P, $Q_3$ | 0.66 | 0.00 | -0.26 | **0.40** |



## Z Score

Why limited ourselves to functional words?
The vocabulary could be different betwen two authors (personal, genre, social, region).

1. *have a bath*, *bike bicycle*, *luncheon*, *sick*, *England*, *Scotch*, *sofa*.
2. *take a bath*, *cycle*, *dinner*, *ill*, *Britain*, *Scottish*, *settee*.

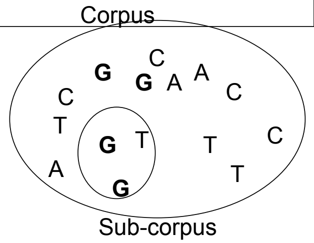
Two authors may use the same words with different intensity, one may over-used a set of forms while the second may under-used them.

Idea: Define the vocabulary specific to an author (genre, type, …) (Ssvoy, 2012)
Variant: see (Pauli & Tuzzi, 2009)

# Z Score

Example

Splitting the whole corpus into two parts.



- Size of the corpus: *n*=15. Subcorpus: 3 (or 1/5 = 0.2)
- Number of **G** in total: 4. In the subcorpus: 2.
- Expected frequency in the subcorpus: 0.2 · 4 = 0.8
- Observed frequency in the sub-corpus: 2
  Thus **G** is over-used (in the subcorpus)!

# Contingency Table

The word "ω" in the sub-corpus and in the rest C-
(C = C- ∪ Sub-corpus)

| | Sub-corpus | C- | C |
|---|---|---|---|
| ω | *a* | *b* | *a*+*b* |
| not "ω" | *c* | *d* | *c*+*d* |
| | *a*+*c* | *b*+*d* | *n* = *a*+*b*+*c*+*d* |

- $n_{\text{sub-corpus}}$ = a + c
- Prob[ω] = (*a*+*b*) / *n*
- Prob[word in Sub-corpus] = (*a*+*c*) / *n*

# Z Score: Example

The word "upon" in Hamilton's papers

| | Hamilton | rest | *Federalist Papers* |
|---|---|---|---|
| "upon" | 370 | 10 | 380 |
| not "upon" | 73,475 | 41,882 | 115,357 |
| | 73,845 | 41,892 | 115,737 |

- Prob[$t_i$] = Prob["upon"] = 380 / 115,737 = 0.003283.
- $n_j$ = 73,845 *a* = 370
- We expect in Hamilton's subcorpus: $n_j \cdot$ Prob[$t_i$] = 242.46
- Z score ("upon" in Hamilton) = 8.2046

# Z Score

- We have a Z score for each term $t_i$ in a subcorpus $D_j$

$$Z\ score(t_{ij}) = \frac{a - (Prob[t_i] \cdot n_j)}{\sqrt{n_j \cdot Prob[t_i] \cdot (1 - Prob[t_i])}}$$

- When comparing two texts, considering all Z scores

$$Dist(D_j, D_k) = \frac{1}{m} \sum_i^m \left(Zscore(t_{ij}) - Zscore(t_{ik})\right)^2$$

## Z Score: Example

The word “on” in Hamilton’s articles

| | Hamilton | rest | *Federalist Papers* |
|---|---|---|---|
| “on” | 374 | 485 | 859 |
| not “on” | 73,471 | 41,407 | 114,878 |
| | 73,845 | 41,892 | 115,737 |

- Prob[$t_i$] = Prob[“on”] = 859 / 115,737 = 0.007422.
- $n_j$ = 73,845 $a$ = 374
- We expect in Hamilton’s subcorpus: $n_j \cdot$ Prob[$t_i$] = 548.08
- Z score (”on" in Hamilton) = -7.46

## Z Score

| | Hamilton | Madison |
|---|---|---|
| Over-used terms | upon | powers |
| | would | confederation |
| | to | department |
| | there | on |
| | courts | congress |
| | kind | and |
| Under-used terms | on | upon |
| | representatives | there |
| | by | would |
| | department | to |

## Intertextual Distance (Labbé, 2007)

- Based on the vocabulary, how can we select part of it?
  Most frequent: Burrows
  Used by every author, every time: Grieve
  Specific vocabulary: Savoy
  Why not all words? Labbé
- The vocabulary choice depends on the subject, genre, epoch, and author
- Define a intertextual distance based on the word types used and their frequencies between two texts.
  But texts with the same genre and epoch.

Labbé C., & Labbé, D. (2001). Intertextual Distance and Authorship Attribution Corneille and Molière. *Journal of Quantitative Linguistic,* 8(3), 213-231.

Labbé, D. (2007). Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, 14(1), 33-80.

## Intertextual Distance (Labbé, 2007)

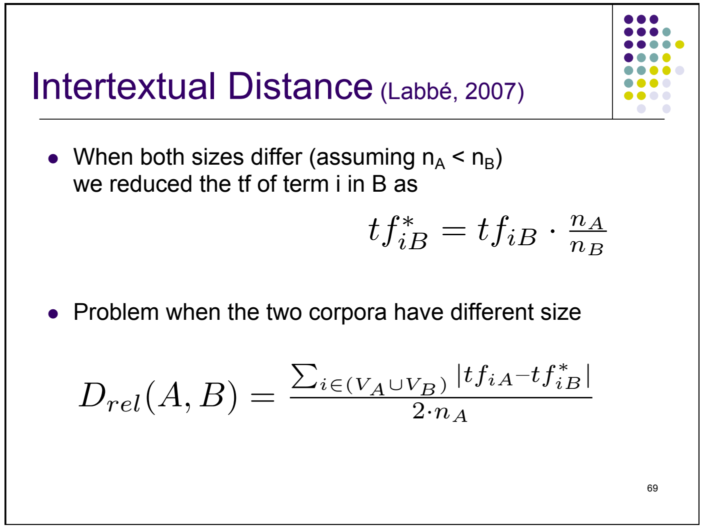
- We define
  $tf_i^A$ = frequency of word type *i* in text A
  $n_A$ = size (number of tokens) of text A
  $V_A$ = vocabulary of text A

$$n_A = \sum_{i \in V_A} tf_{iA}$$

- Distance D(A,B) between Text A and Text B (similar size)

$$D(A,B) = \sum_{i \in (V_A \cup V_B)} |tf_{iA} - tf_{iB}| \quad with\ n_A = n_B$$

- D(A,B) = 0
  both texts use the same words with the same frequencies
- Otherwise D(A,B) > 0 (lim: $n_A + n_B$)
  the number of tokens that differ



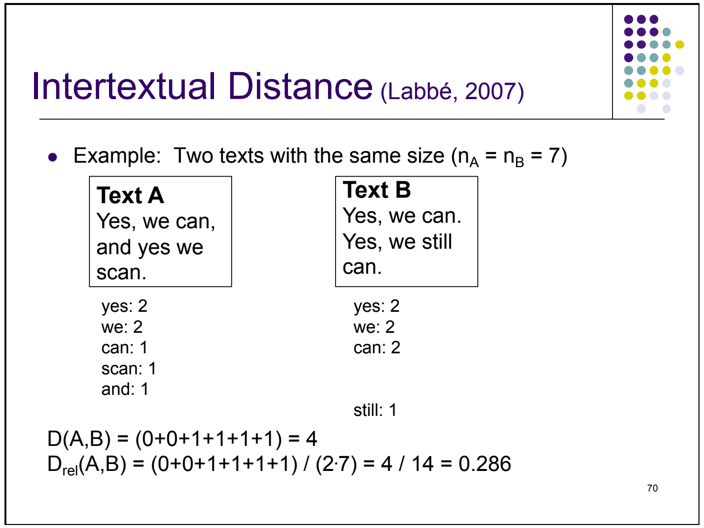
## Intertextual Distance (Labbé, 2007)

- When both sizes differ (assuming $n_A < n_B$) we reduced the tf of term i in B as

$$tf^*_{iB} = tf_{iB} \cdot \frac{n_A}{n_B}$$

- Problem when the two corpora have different size

$$D_{rel}(A,B) = \frac{\sum_{i \in (V_A \cup V_B)} |tf_{iA} - tf^*_{iB}|}{2 \cdot n_A}$$



## Intertextual Distance (Labbé, 2007)

- Example: Two texts with the same size ($n_A = n_B = 7$)

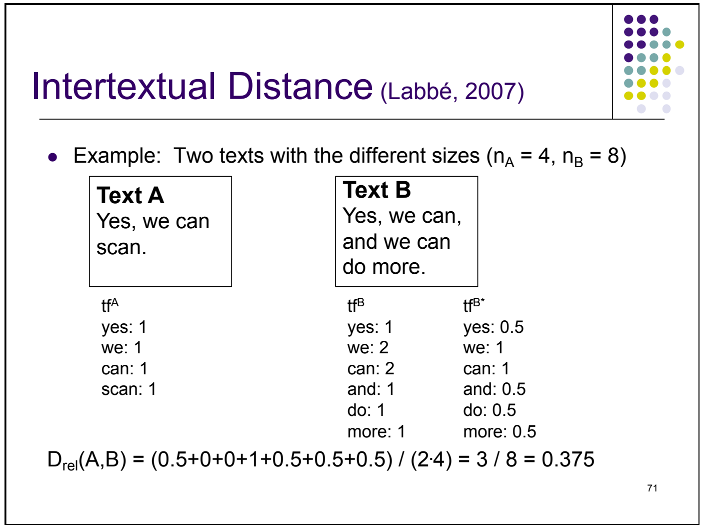
**Text A**
Yes, we can, and yes we scan.

**Text B**
Yes, we can. Yes, we still can.

| | |
|---|---|
| yes: 2 | yes: 2 |
| we: 2 | we: 2 |
| can: 1 | can: 2 |
| scan: 1 | |
| and: 1 | |
| | still: 1 |

D(A,B) = (0+0+1+1+1+1) = 4

$D_{rel}$(A,B) = (0+0+1+1+1+1) / (2·7) = 4 / 14 = 0.286



## Intertextual Distance (Labbé, 2007)

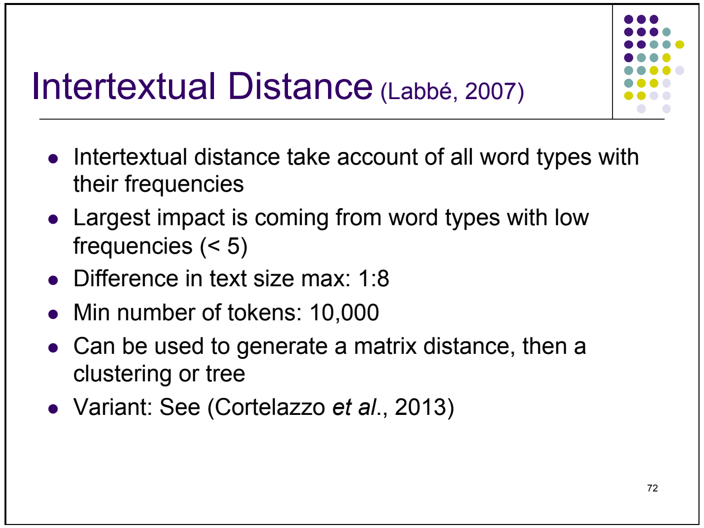
- Example: Two texts with the different sizes ($n_A = 4$, $n_B = 8$)

**Text A**
Yes, we can scan.

**Text B**
Yes, we can, and we can do more.

| $tf^A$ | $tf^B$ | $tf^{B^*}$ |
|---|---|---|
| yes: 1 | yes: 1 | yes: 0.5 |
| we: 1 | we: 2 | we: 1 |
| can: 1 | can: 2 | can: 1 |
| scan: 1 | and: 1 | and: 0.5 |
| | do: 1 | do: 0.5 |
| | more: 1 | more: 0.5 |

$D_{rel}$(A,B) = (0.5+0+0+1+0.5+0.5+0.5) / (2·4) = 3 / 8 = 0.375

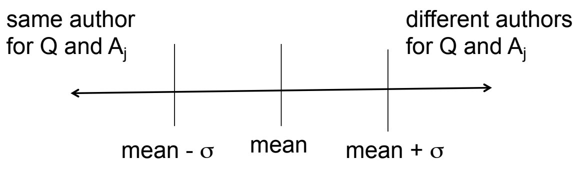


## Intertextual Distance (Labbé, 2007)

- Intertextual distance take account of all word types with their frequencies
- Largest impact is coming from word types with low frequencies (< 5)
- Difference in text size max: 1:8
- Min number of tokens: 10,000
- Can be used to generate a matrix distance, then a clustering or tree
- Variant: See (Cortelazzo *et al.*, 2013)

## Intertextual Distance (Labbé, 2007)

- Decision according to the $D_{rel}(Q,A_j)$


same author for Q and $A_j$

different authors for Q and $A_j$

mean - σ

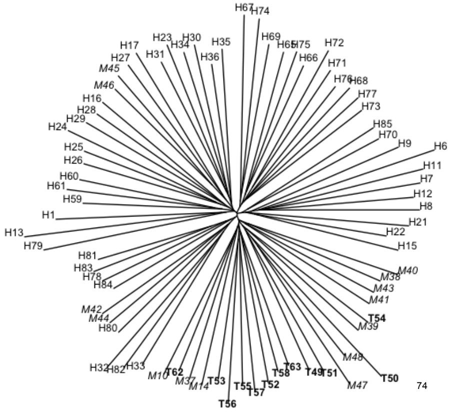
mean

mean + σ

## Intertextual Distance (Labbé, 2007)

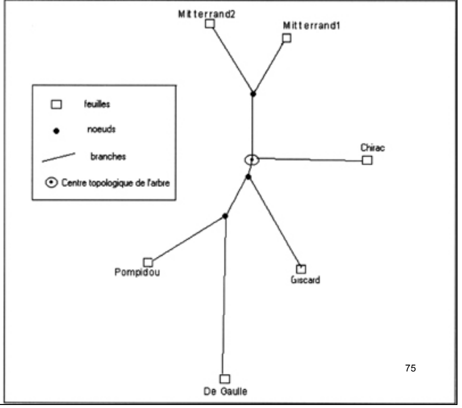
Federalist Papers

$D_{rel}(A,B)$ & Clustering



## Intertextual Distance (Labbé, 2007)

Clustering & $D_{rel}(A,B)$

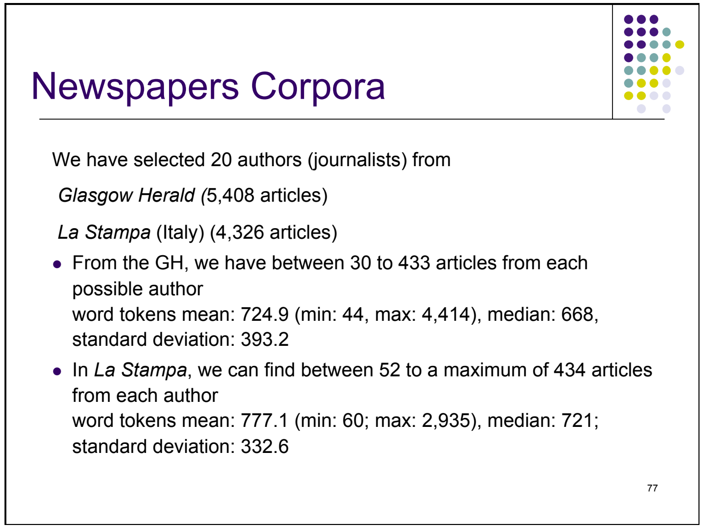


## Newspapers Corpora

*Glasgow Herald* (1995)

*La Stampa* (Italy) (1994)

LA STAMPA.it
EDIZIONI LOCALI: TORINO - CUNEO - AOSTA - ASTI - NOVARA - VCO - VERCELLI - BIELLA | PROSSIMA USCITA: ALESSANDRIA
ATTUALITÀ
OPINIONI
ECONOMIA
SPORT
TORINO
CULTURA
SPETTACOLI
MOTORI
DONNA
CUCINA
SALUTE
VIAGGI
EXTR@
SPECIALI
HOME | POLITICA | ESTERI | CRONACHE | COSTUME | TECNOLOGIA | SCIENZA | AMBIENTE | LAZAMPA | I TUOI DIRITTI | MULTIMEDIA



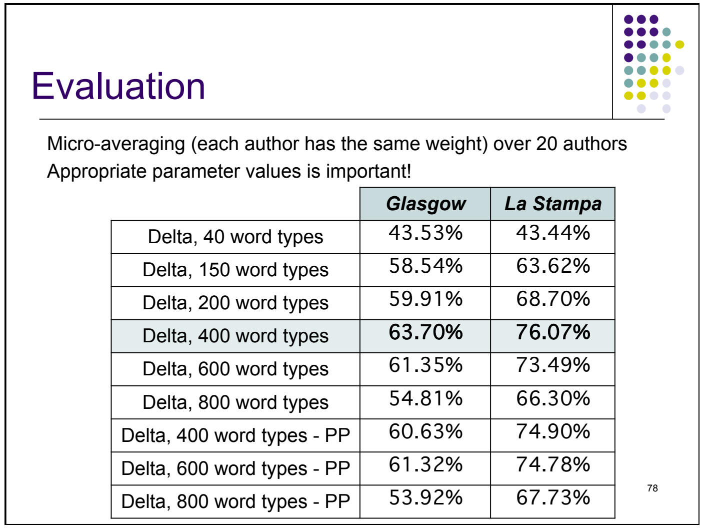
# Newspapers Corpora

We have selected 20 authors (journalists) from

*Glasgow Herald* (5,408 articles)

*La Stampa* (Italy) (4,326 articles)

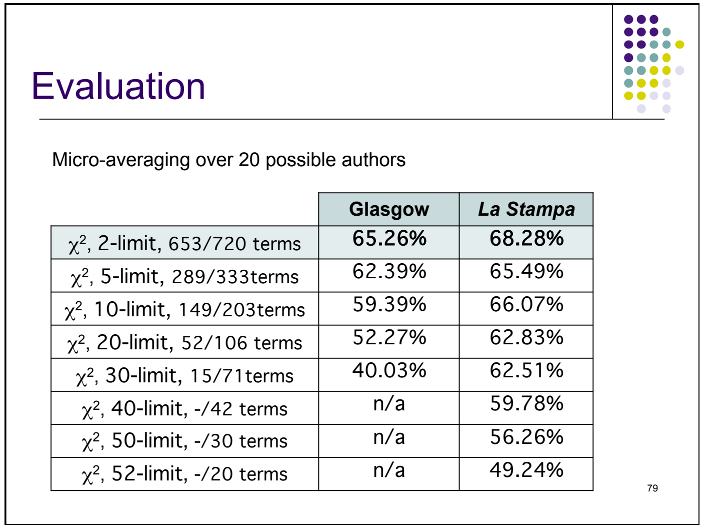
- From the GH, we have between 30 to 433 articles from each possible author
  word tokens mean: 724.9 (min: 44, max: 4,414), median: 668, standard deviation: 393.2
- In *La Stampa*, we can find between 52 to a maximum of 434 articles from each author
  word tokens mean: 777.1 (min: 60; max: 2,935), median: 721; standard deviation: 332.6



# Evaluation

Micro-averaging (each author has the same weight) over 20 authors
Appropriate parameter values is important!

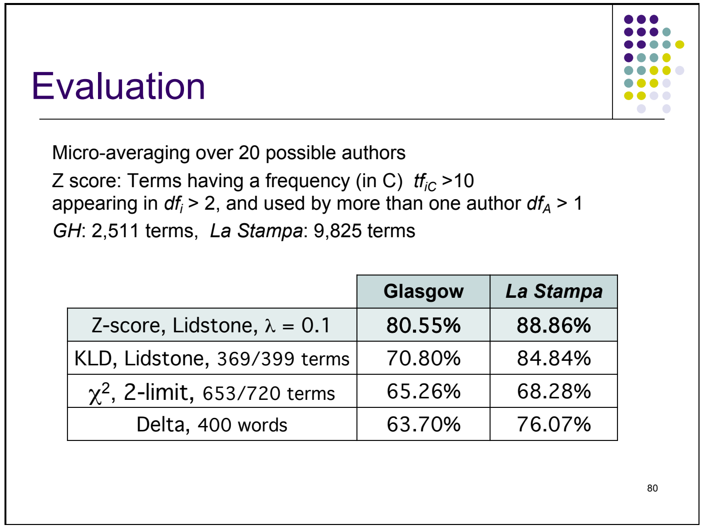
| | *Glasgow* | *La Stampa* |
|---|---|---|
| Delta, 40 word types | 43.53% | 43.44% |
| Delta, 150 word types | 58.54% | 63.62% |
| Delta, 200 word types | 59.91% | 68.70% |
| Delta, 400 word types | **63.70%** | **76.07%** |
| Delta, 600 word types | 61.35% | 73.49% |
| Delta, 800 word types | 54.81% | 66.30% |
| Delta, 400 word types - PP | 60.63% | 74.90% |
| Delta, 600 word types - PP | 61.32% | 74.78% |
| Delta, 800 word types - PP | 53.92% | 67.73% |



# Evaluation

Micro-averaging over 20 possible authors

| | Glasgow | *La Stampa* |
|---|---|---|
| $\chi^2$, 2-limit, 653/720 terms | **65.26%** | **68.28%** |
| $\chi^2$, 5-limit, 289/333terms | 62.39% | 65.49% |
| $\chi^2$, 10-limit, 149/203terms | 59.39% | 66.07% |
| $\chi^2$, 20-limit, 52/106 terms | 52.27% | 62.83% |
| $\chi^2$, 30-limit, 15/71terms | 40.03% | 62.51% |
| $\chi^2$, 40-limit, -/42 terms | n/a | 59.78% |
| $\chi^2$, 50-limit, -/30 terms | n/a | 56.26% |
| $\chi^2$, 52-limit, -/20 terms | n/a | 49.24% |



# Evaluation

Micro-averaging over 20 possible authors
Z score: Terms having a frequency (in C) $tf_{iC}$ >10
appearing in $df_i$ > 2, and used by more than one author $df_A$ > 1
*GH*: 2,511 terms, *La Stampa*: 9,825 terms

| | Glasgow | *La Stampa* |
|---|---|---|
| Z-score, Lidstone, $\lambda = 0.1$ | **80.55%** | **88.86%** |
| KLD, Lidstone, 369/399 terms | 70.80% | 84.84% |
| $\chi^2$, 2-limit, 653/720 terms | 65.26% | 68.28% |
| Delta, 400 words | 63.70% | 76.07% |

## Hidden Questions / Problems

- Split clearly between a training set and a test set
- Each model has its own limits
- Size of the (disputed / training) texts
  100 tokens to 10,000 tokens
  Better performance
  with long texts, long profiles, few authors
- (Un)Balanced set in generating the author's profiles
- Type of text (e.g., dialogue, descriptive, narrative)
- The style may change during the author's life
- Style related to a given character (or set of characters) for a given author

- Der Teufel liegt im Detail

## Conclusion

- Authorship attribution
  - The result of computational linguistics are always matters of probability, not certainty. …After all, we are dealing with writers who are at liberty to imitate each others, to try new styles, and to write differently for a particular occasion or in a new genre, …"
    (Craig & Kinney, 2009, p. 24-25)

- **L'Aquila quake: Italian scientists guilty**

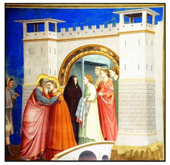
- Explain the decision with stylistic elements

## Conclusion

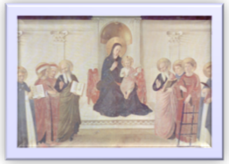
- Next steps
  - Consider other representation than isolated words
    *n*-gram of characters, *n*-gram of words,
    POS, *n*-gram of POS
  - Other languages
  - Other paradigm (machine learning) to promote better classifier(?)
  - Author profiling
  - Other medium

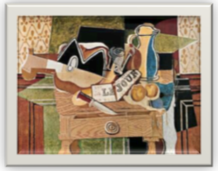
## Being in Padova…

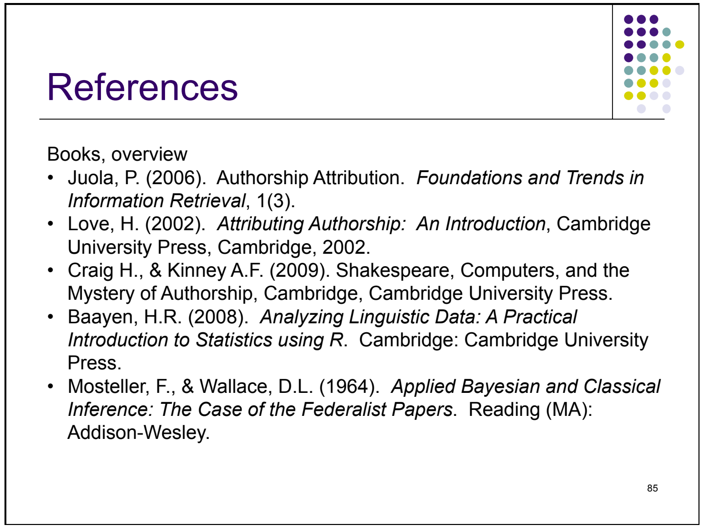
Use your eyes…
Knowing the author of this three paintings,



Are the last two painted by the same author?



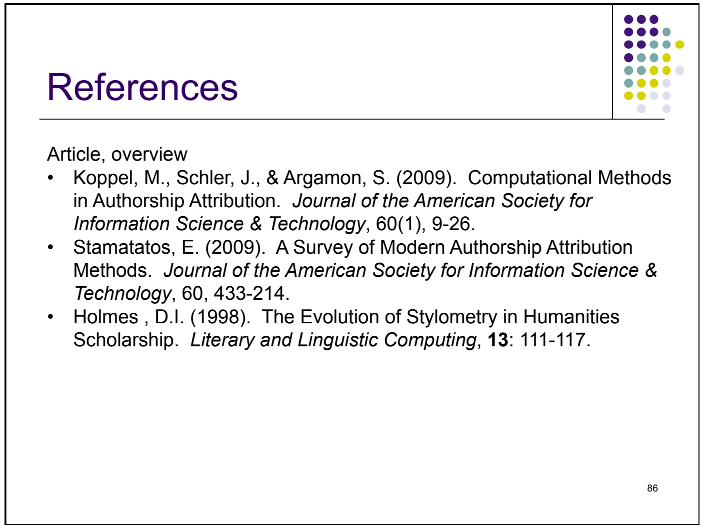




## References

Books, overview

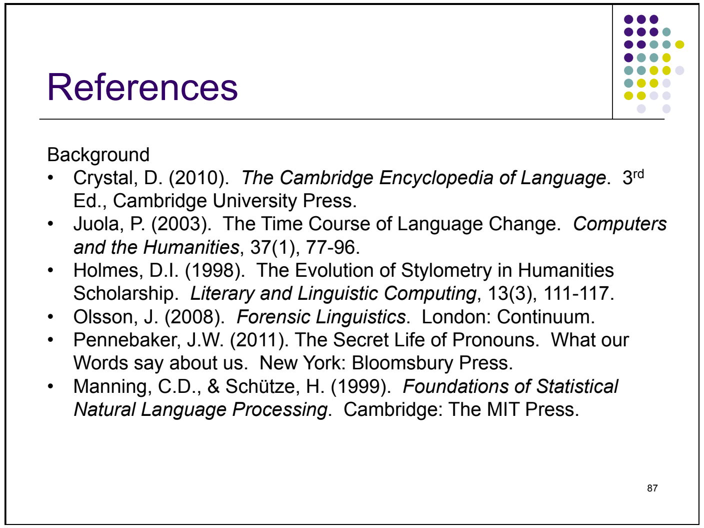
- Juola, P. (2006). Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3).
- Love, H. (2002). *Attributing Authorship: An Introduction*, Cambridge University Press, Cambridge, 2002.
- Craig H., & Kinney A.F. (2009). Shakespeare, Computers, and the Mystery of Authorship, Cambridge, Cambridge University Press.
- Baayen, H.R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.
- Mosteller, F., & Wallace, D.L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Reading (MA): Addison-Wesley.



## References

Article, overview

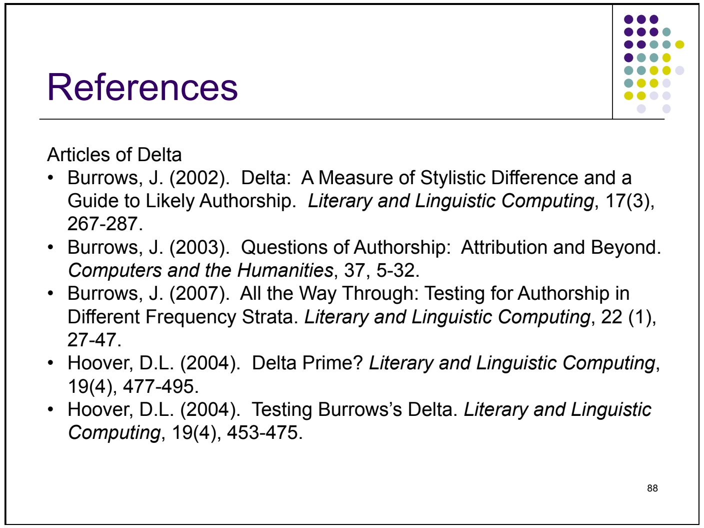
- Koppel, M., Schler, J., & Argamon, S. (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science & Technology*, 60(1), 9-26.
- Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science & Technology*, 60, 433-214.
- Holmes , D.I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, **13**: 111-117.



## References

Background

- Crystal, D. (2010). *The Cambridge Encyclopedia of Language*. 3rd Ed., Cambridge University Press.
- Juola, P. (2003). The Time Course of Language Change. *Computers and the Humanities*, 37(1), 77-96.
- Holmes, D.I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3), 111-117.
- Olsson, J. (2008). *Forensic Linguistics*. London: Continuum.
- Pennebaker, J.W. (2011). The Secret Life of Pronouns. What our Words say about us. New York: Bloomsbury Press.
- Manning, C.D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.



## References

Articles of Delta

- Burrows, J. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3), 267-287.
- Burrows, J. (2003). Questions of Authorship: Attribution and Beyond. *Computers and the Humanities*, 37, 5-32.
- Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22 (1), 27-47.
- Hoover, D.L. (2004). Delta Prime? *Literary and Linguistic Computing*, 19(4), 477-495.
- Hoover, D.L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4), 453-475.

# References

Articles of various approaches

- Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22, 251-270.
- Pauli, F., & Tuzzi, A. (2009). The end of year addresses of the Presidents of the Italian Republic (1948–2006): discoursal similarities and differences. Glottometrics, 18, 40–51.
- Savoy, J. (2012). Authorship Attribution Based on Specific Vocabulary. *ACM Transaction in Information Systems*, 30(2).
- Savoy, J. (2013). Comparative Evaluation of Term Selection Functions for Authorship Attribution. *Literary and Linguistic Computing*, (to appear).
- Zhao, Y., & Zobel, J. (2007). Entropy-Based Authorship Search in Large Document Collection. In *Proceedings ECIR* (pp. 381-392). Berlin: Springer, LNCS #4425.

# References

Articles on Labbé's methods

- Cortelazzo, M.A., Nadalutti, P. & Tuzzi, A. (2013). Improving Labbé's Intertextutal Distance: Testing a Revised Version on a Large Corpus of Italian Literature. *Journal of Quantitative Linguistics.* 20(2), 125-152.
- Labbé C., & Labbé, D. (2001). Intertextual Distance and Authorship Attribution Corneille and Molière. *Journal of Quantitative Linguistic,* 8 (3), 213-231.
- Labbé, D. (2007). Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, 14 (1), 33-80.
- Labbé C., & Labbé, D. (2007). A Tool for Literary Studies: Intertextual Distance and Tree Classification. *Literacy & Linguistic Computing,* 381-392.

# References

Other articles

- Binongo, J.N.G., & Smith, M.W. (1999). The Application of Principal Component Analysis to Stylometry. *Literary and Linguistic Computing.* 14(4), 445-465.
- Cilibrasi, R., Vitanyi, M.B. (2004). Clustering by Compression. *IEEE Transactions on Information Theory*, 51(4), 1523-1545.
- Miranda Garcia, A., & Calle Martin, J. (2007). Function Words in Authorship Attribution Studies. *Literary & Linguistic Computing*, 22(1), 49-66.
- Hoover, D.L. (2003). Another Perspective on Vocabulary Richness. *Computers and the Humanities.* 37, 151-178.
- Koppel, M., Schler, J., & Argamon, S. (2007). Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research*, 8, 1261-1276.
- Savoy, J. (2013). Comparative Evaluation of Term Selection Functions for Authorship Attribution. *Literary & Linguistic Computing*.

# Forensic Linguistics

DEAR BILL,

I SUPPOSE YOU THOUGHT I WOULD FORGET BUT YOU ARE WRONG HOW COULD I FORGET A RAT LIKE YOU. I HAVE SENT A LETTER WITH ALL YOUR PAST DETAILS TO THE PRESIDENT. ALL YOUR DEBTS AND PAST MISSDEMEANOURS. IF YOU DON'T RESIGN FROM THE COUNCIL IMMEDIATELY THE PRESS WILL PRINT A LIST OF ALL YOUR DEBTS BOTH LOCALLY AND NATIONALLY… YOU MIGHT BE ABLE TO FOOL SOME PEOPLE BUT NOT ME. YOUR FORGET I HAVE KNOWN YOU FOR ALL OF YOUR LIFE.

## Admission in Court (US)

1. Knowledge and stature: the exert must have sufficient knowledge of the subject.
2. Testing: the technique must be empirically tested.
3. Peer review: subjected to a peer review.
4. Scientific method: the error rate is known
5. Straightforwardness: the technique can be explained with clarity and simplicity.

Forensic Linguistic (Olssen, 2008)

## Example with the *Federalist*

- Spelling variation
  *while* (Hamilton) vs. *whilst* (Madison)
- In the vocabulary used only by one:
  Hamilton: destruction, offensive, defensive, contribute
  Madison: violence, fortune, although
- Vocabulary used more frequently by one
  *considerable* (13 Hamilton, 4 Madison)
  *voice* (1, 8)
  *again* (1,7)
  *language* (2,10)

## Evaluation: Federalist Papers

12 disputed papers assigned to Madison

| | ***Default*** | ***Error*** |
|---|---|---|
| Delta, 40 word types | 10 / 12 | #55, #56 |
| Delta, 50 word types | 9 / 12 | #55, #56, #63 |
| Delta, 100 word types | 10 / 12 | #55, #56 |
| Delta, 150 word types | 11 / 12 | #56 |
| Delta, 200 word types | 9 / 12 | #50, #56, #57 |
| KLD, Zhao | 9 / 12 | #49, #55, #57 |
| KLD, Hughes | 12 / 12 | |
| Intertextual distance | No assign. | |
| Intertextual & Clustering | 12 / 12 | |

## Z Score: Example

The word “Bush” in McCain’s speeches in 2008 (= $D_j$)
vs. all other US electoral speeches

| | **McCain’08** | **rest** | **C** |
|---|---|---|---|
| “Bush” | 26 | 398 | 424 |
| not “Bush” | 154,339 | 474,331 | 628,670 |
| | 154,365 | 474,729 | 629,094 |

- Prob[$t_i$] = Prob[“Bush” in C] = 424 / 629,094 = 0.000674.
- $n_j$ = 154,365 $a$ = 26
- We expect in McCain'08 (= $D_j$): $n_j \cdot$Prob[$t_i$] = 104.04
- Z score ("Bush" in McCain'08) = -7.65

## I and Obama

« I » (and me) is the prototypical stealth word.

When a person uses a lot of « I »: arrogant, self-confident

G. Will (*Washington Post*, June 7, 2009)

S. Fish (*New York Times*, June 7, 2009)

pointed out the Obama's frequent use of « I »

(Pennebacker, 2011)

| ***State of the Union*** | **Obama** | **Bush** | **Clinton** |
|---|---|---|---|
| "I" | 1.13% | 0.62% | 5.76% |
| "we" | 3.73% | 2.81% | 17.21% |

## Psychological Profile

We can establish the psychological profile of the writer according to four dimensions (MBTI indicator):

1. Extroversion vs. Introversion
   (social interaction vs. solitude)
2. Intuition – Sensing
   (prefers theoretical info vs. perceiving the info)
3. Thinking – Feeling
   (logical decision vs. decision according to subjective values)
4. Judgment – Perception
   (judgement accroding to my perceptions vs. don't quickly jump to a conclusion)

(Noecker, Ryan, Juola, *LLC*, 2013), (Pennebacker, 2011)

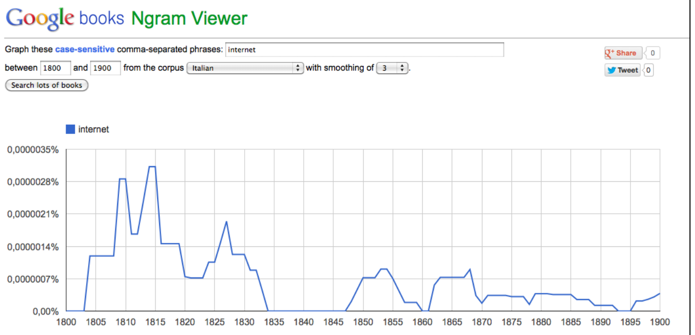
## Age

- Older people tend to use more future tense
- Young people tend to use more past tense

- Upper class: more « we »
- Lower class: more « I »

(Pennebacker, 2011)

## Why data quality matters

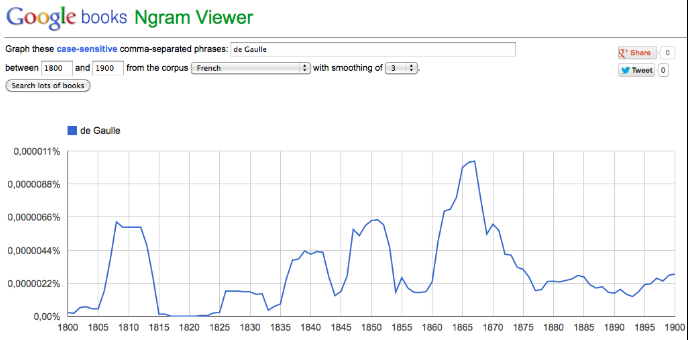


Google books Ngram Viewer

Graph these case-sensitive comma-separated phrases: internet

between 1800 and 1900 from the corpus Italian with smoothing of 3.

# Why data quality matters



Google books Ngram Viewer

Graph these case-sensitive comma-separated phrases: de Gaulle

between 1800 and 1900 from the corpus French with smoothing of 3.

# Zeta: Less Frequently Used Words

- Instead on focussing on very frequent words, focus on words used more frequently by a given author.
  E.g., Shakespeare uses more *gentle*, *answer* but less frequently *brave*, *sure*, *hopes*, or *beseech*.
- Split the texts into blocks (20,000 tokens), form a set of texts written by A, and a counter-set written by others (-A)
- Select word types having df$^A$ ≥ δ (e.g., in 3 blocks) (relatively frequent in blocks written by A)
  and word types must have df$^{-A}$ ≥ δ (e.g., 3)
- Binary view: term present or not

Craig H., Kinney A.F. (2009). Shakespeare, Computers, and the Mystery of Authorship, Cambridge, Cambridge University Press.
Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22 (1), 27-47.

# Zeta: Less Frequently Used Words

- First solution: for a given term $t_i$, we can count:
  - the number of texts (blocks) in which the term $t_i$ appears (or df$_i$)
  - the number of texts (blocks) where it does not appear (or df$_{-i}$)
  - the ratio (df$_i$ / df$_{-i}$)
- But we can include the fact that the author was A or B (df$_i^A$ number of blocks written by A with term t$_i$
  |T$^A$| denotes the number of texts written by A)

$$index(t_i, A, B) = \frac{|df_i^A|}{|T^A|} + \frac{|df_{-i}^B|}{|T^B|}$$

Craig H., Kinney A.F.(2009) Shakespeare, Computers, and the Mystery of Authorship, Cambridge, Cambridge University Press.

# Zeta: Less Frequently Used Words

$$index(t_i, A, B) = \frac{|df_i^A|}{|T^A|} + \frac{|df_{-i}^B|}{|T^B|}$$

- If a term $t_i$ appears in all and only in texts written by A, the index will be 1 + 1 = 2
- If a term $t_i$ is used by both writers, and in all of their texts, the index will be 1 + 0 = 1
- If a term $t_i$ is used by both writers in the same proportion (e.g., 30%), the index will be 0.3 + 0.7 = 1
- If a term $t_i$ is used only by B (with a proportion of 20%), the index will be 0 + 0.8 = 0.8

Craig H., Kinney A.F.(2009) Shakespeare, Computers, and the Mystery of Authorship, Cambridge, Cambridge University Press.

# Lexical Test

Shakespeare

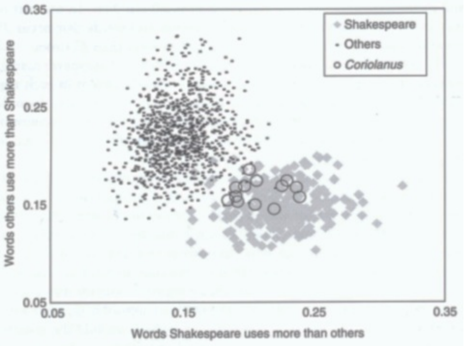


Figure 2.2 Lexical-words test: 2000-word Shakespeare segments versus 2000-word segments by others, with 2000-word segments of *Coriolanus*.

# Lexical Test

Shakespeare

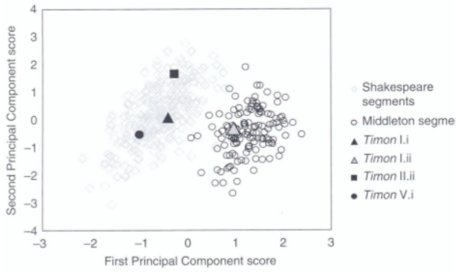
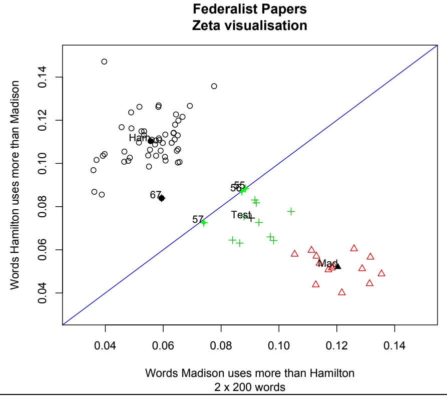
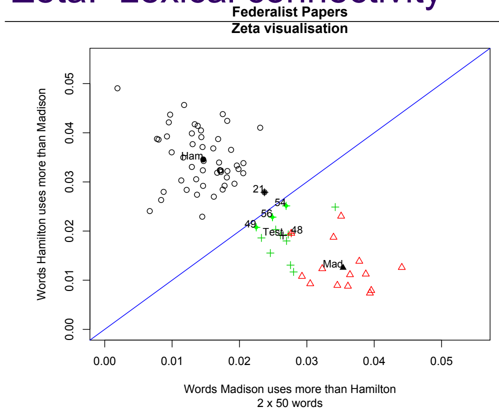


Figure 2.5 Function-words test: 2000-word Shakespeare segments, 2000-word Middleton segments, and 4 scenes from *Timon of Athens*.

# Zeta: Lexical connectivity



# Zeta: Lexical connectivity


Federalist Papers
Zeta visualisation
Words Hamilton uses more than Madison
Words Madison uses more than Hamilton
2 x 50 words
Ham
Test
Mad