Grieve 2007: Quantitative Authorship Attribution: An Evaluation of Techniques

Zarah Weiß

November 18th, 2015

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Introduction

Textual Measurements
- Length Measures
- Vocabulary Richness Measures
- Frequency Measures

The Algorithm

The Corpus

Experiment & Results
- Experiment
- Results
- Combination of Techniques

Conclusion

References

Discussion

Introduction

Quantitative Authorship Attribution

- Determine author from set of possible authors
- Based on corpus of author set
- Based on textual measures (features)
- Attribution algorithm compares anonymous text with known author data
- Mendenhall (1887) on Shakespeare plays

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Introduction

Grieve 2007

- Overview over 39 most common features for authorship attribution
- First comprehensive feature set evaluation
- Uses identical data set
- Uses identical attribution algorithm
- Proposes more accurate approach combining promising features

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Textual Measurements

Length Measures

	Word-Length	Sentence-Length
Average length	$\frac{\#\ digits\ +\ \#\ graphemes}{\#\ "words"}$!	$\frac{(\#\ "words"\ \mid\ \#\ characters!)}{\#\ sentences}$
Distribution rel. freq.	$\frac{\#\ "words"\ of\ length\ n}{\#\ "words"}$	$\frac{\#\ sentences\ of\ length\ n}{\#\ sentences}$

Table: Length measures evaluated in Grieve 2007.

- For $n = 1, \ldots, N$ (for varying N)
- For sentence frequency distribution in characters n as range, e.g. 1 to 10 characters
- With sentence length being measured in
  1. # "words"
  2. # characters

Grieve 2007: Quantitative Authorship Attribution: An Evaluation of Techniques

Zarah Weiß

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Textual Measurements

Length Measures

	Word-Length	Sentence-Length
Average length	$\frac{\#\ digits\ +\ \#\ graphemes}{\#\ "words"}$!	$\frac{(\#\ "words"\ \mid\ \#\ characters!)}{\#\ sentences}$
Distribution rel. freq.	$\frac{\#\ "words"\ of\ length\ n}{\#\ "words"}$	$\frac{\#\ sentences\ of\ length\ n}{\#\ sentences}$

Table: Length measures evaluated in Grieve 2007.

- For $n = 1, \ldots, N$ (for varying N)
- For sentence frequency distribution in characters n as range, e.g. 1 to 10 characters
- With sentence length being measured in
  1. # "words"
  2. # characters
- length("Chris drank an espresso .") = ?

Grieve 2007: Quantitative Authorship Attribution: An Evaluation of Techniques

Zarah Weiß

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Textual Measurements

Length Measures

	Word-Length	Sentence-Length
Average length	$\frac{\#\ digits\ +\ \#\ graphemes}{\#\ "words"}$!	$\frac{(\#\ "words"\ \mid\ \#\ characters!)}{\#\ sentences}$
Distribution rel. freq.	$\frac{\#\ "words"\ of\ length\ n}{\#\ "words"}$	$\frac{\#\ sentences\ of\ length\ n}{\#\ sentences}$

Table: Length measures evaluated in Grieve 2007.

- For $n = 1, \ldots, N$ (for varying N)
- For sentence frequency distribution in characters n as range, e.g. 1 to 10 characters
- With sentence length being measured in
  1. # "words"
  2. # characters
- length("Chris drank an espresso .") = ?
  1. 4 (dot is neither grapheme nor digit)

Textual Measurements

Length Measures

	Word-Length	Sentence-Length
Average length	$\frac{\#\ digits\ +\ \#\ graphemes}{\#\ "words"}$!	$\frac{(\#\ "words"\ \mid\ \#\ characters!)}{\#\ sentences}$
Distribution rel. freq.	$\frac{\#\ "words"\ of\ length\ n}{\#\ "words"}$	$\frac{\#\ sentences\ of\ length\ n}{\#\ sentences}$

Table: Length measures evaluated in Grieve 2007.

- For $n = 1, \ldots, N$ (for varying N)
- For sentence frequency distribution in characters n as range, e.g. 1 to 10 characters
- With sentence length being measured in
  1. # "words"
  2. # characters
- length("Chris drank an espresso .") = ?
  1. 4 (dot is neither grapheme nor digit)
  2. 25 (again, no dot)

Textual Measurements

Vocabulary Richness Measures

Unrestricted type-"word" ratio: $\frac{\#\ types}{\#\ "words"}$

- Issue?

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Textual Measurements

Vocabulary Richness Measures

Unrestricted type-"word" ratio: $\frac{\#\ types}{\#\ "words"}$

- Issue? Sensitive to text length!

Type Token Ratio variations:

- Guiraud's R: $\frac{\#\ types}{\sqrt{\#\ "words"}}$
- Herdan's C: $\frac{log(\#\ types)}{log(\#\ "words")}$
- Dugat's k: $\frac{log(\#\ types)}{log(log(\#\ "words"))}$
- Tuldava's LN: $\frac{1\ -\ (\#\ types)^2}{(\#\ types)^2 \times log(\#\ "words"))}$
- Restricted type-"word" ratio: $\frac{\#\ first\ \ n\ \ types}{\#\ first\ \ n\ "words"}$, with n being $\#$ "*words*" in shortest writing sample

Textual Measurements

Vocabulary Richness Measures

Type Token Ratio variations:

- Sichel's S and Michéa's M: $\frac{\#\ types\ occurring\ 2\ times}{\#\ tokens}$
- Honoré's H: $\frac{100 \times log(\#\ "words")}{(1\ -\ \#\ types\ occurring\ 1\ time)/\#\ types}$
- Yule's K and Simpson's D: $10^4 \times \frac{\sum i^2 \times \#\ types\ occurring\ i\ times\ -\ \#\ "words"}{(\#\ "words")^2}$

Other lexical diversity measures:

- Entropy: $-100 \times \sum_v p_v \times log(p_v)$,
  with p_v = relative frequency of v^{th} most frequent type
- W: $(\#\ "words")^{\#\ types\ -\ a}$, with some constant a

For evaluation of LD measures, see McCarthy & Jarvis (2007, 2010)!

Textual Measurements

Grapheme Frequency

Simple grapheme profile[1]: $\frac{\#\ instances\ of\ grapheme\ i}{\#\ graphemes}$

- For each $i \in$ set(alphabet)

Single-position grapheme profile: $\frac{\#\ instances\ of\ grapheme\ i\ in\ position\ p}{\#\ "words"\ containing\ position\ p}$

- For each $i \in$ set(alphabet)
- For varying positions p within a "word" (first, second, . . . , last grapheme)

[1] All profiles are frequency distributions! I.e. one profile per text!

Textual Measurements

Grapheme Frequency

Word-internal grapheme profile[2]: $\frac{\#\ \text{"}words\text{"}\ \ containing\ \ grapheme\ \ i}{\#\ \text{"}words\text{"}}$

- For each $i \in$ set(alphabet)

Multi-position grapheme profile: $\frac{\#\ instances\ \ of\ \ I_p^P}{\#\ \text{"}words\text{"}\ \ containing\ \ positions\ \ [p{:}(p{+}n)]}$

- With I being a number of graphemes at positions p to P (not necessarily adjacent)
- I.e. multiple single-position grapheme profiles
- For varying positions p within a "word" (e.g. first and last 3 graphemes in a "word")

[2] All profiles are frequency distributions! I.e. one profile per text!

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Textual Measurements

Word Frequency & Positional Stylometry

Simple word profile[3]: $\frac{\#\ instances\ "word"\ t}{\#\ "words"}$

- For each $t \in$ set(high frequency words)
- With varying minimum frequency cut off for set(high frequency words)

Single-position word profile: $\frac{\#\ instances\ of\ "word"\ t\ in\ postion\ p}{\#\ sentences\ containing\ position\ p}$

- For each "word" t in the text
- With varying positions p in a sentence (first, second, . . . , last "word")

[3] All profiles are frequency distributions! I.e. one profile per text!

Multi-position word profile[4]: $\frac{\#\ instances\ of\ I_p^{p+n}}{\#\ sentences\ containing\ position\ [p{:}(p+n)]}$

- With I being a "word" sequence of length n + 1 starting at position p
- I.e. multiple single-position word profiles
- For varying positions p within a sentence (e.g. first 3 "words" in a sentence)

[4] All profiles are frequency distributions! I.e. one profile per text!

Grieve 2007: Quantitative Authorship Attribution: An Evaluation of Techniques

Zarah Weiß

Introduction

Textual Measurements

Length Measures

Vocabulary Richness Measures

Frequency Measures

The Algorithm

The Corpus

Experiment & Results

Experiment

Results

Combination of Techniques

Conclusion

References

Discussion

Textual Measurements

Punctuation Mark Frequency

Simple punctuation mark profile[5]: $\frac{\#\ punctuation\ mark\ m}{[\#\ characters \mid \#\ punctuation\ marks \mid \#\ "words"\ !]}$

- With $m \in$ set(punctuation marks) $= \{.\ ,\ :\ ;\ -\ ?\ (\ '\}$!

Punctuation and grapheme profile: $\frac{\#\ instances\ of\ character\ i}{\#\ graphemes + \#\ punctuation\ marks}$?

- For each $i \in$ set(alphabet) $\cup$ set(punctuation marks)

Punctuation and word profile: $\frac{\#\ instances\ of\ string\ t}{\#\ "words" + \#\ punctuation\ marks}$?

- For each $t \in$ set("words") $\cup$ set(punctuation marks)

[5] All profiles are frequency distributions! I.e. one profile per text!

Textual Measurements

Collocation Frequency

N-gram profile[6]: $\frac{\#\ character\ \ n-gram\ \ g}{\#\ character\ \ n-grams}$

- With $g \in$ set(high frequency character n-grams)
- Overall eight profiles for $2 \leq n \leq 9$
- With varying minimum frequency cut off for set(high frequency character n-grams)
- Character-Level N-Gram Frequency!

[6] All profiles are frequency distributions! I.e. one profile per text!

Textual Measurements

Collocation Frequency

N-word collocation profile[7]: $\frac{\#\ \text{"}word\text{"}\ \ n-gram\ \ g}{\#\ \text{"}word\text{"}\ \ n-grams}$

- With $g \in$ set(highly frequency "word" n-grams), i.e. collocations
- Overall two profiles for $2 \leq n \leq 3$
- With varying minimum frequency cut off for set(highly frequency "word" bigrams)
- "word"-Level N-Gram Frequency!

[7] All profiles are frequency distributions! I.e. one profile per text!

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

The Algorithm

The Workflow

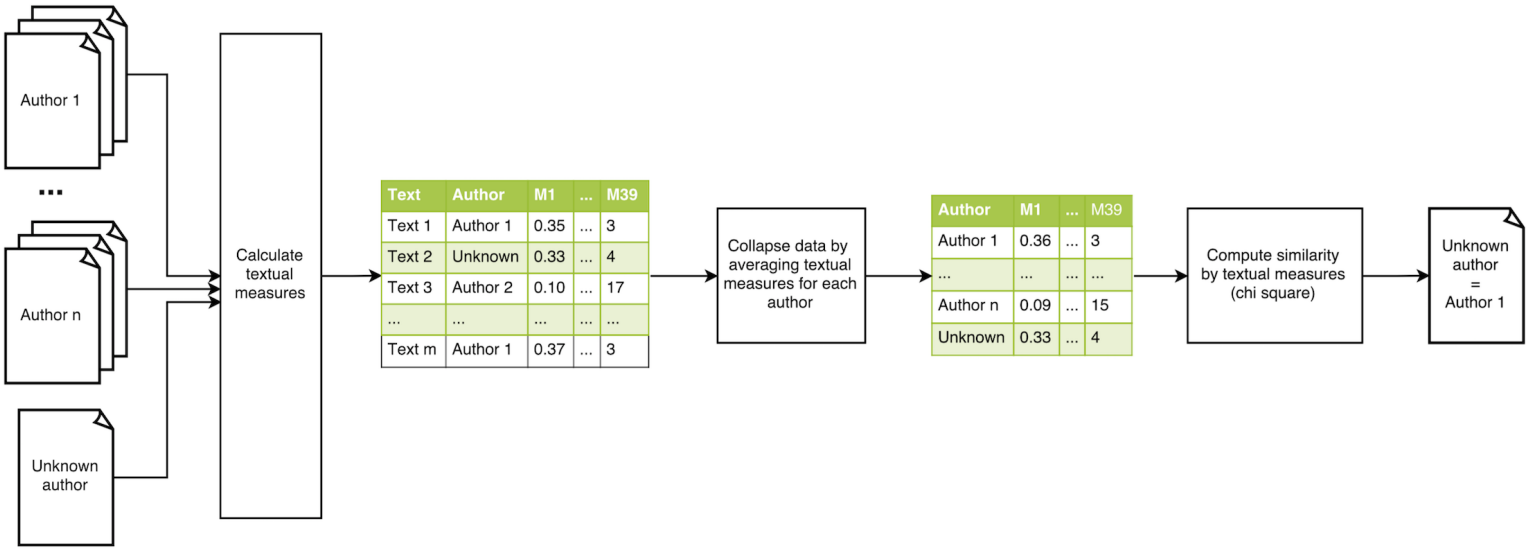

Figure: Workflow of the (generalized) attribution algorithm.

The Algorithm

Statistics

- Similarity of authors measured with chi-square test
- Most common statistic for authorship attribution
- Measures dependence / independence of properties given their frequencies
- Question: Could the sample have been drawn from the population?

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

The Algorithm

Statistics

Chi-square: $\chi^2 = \sum_i^r \sum_j^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

- With O being observed frequencies of a sample (unknown author's profile)
- With E being expected frequencies of a population (other authors' profile)
- Grieve 2007 tests each textual measure profile separately!

Expected frequency (E_{ij}): $\frac{O_{i.} \times O_{.j}}{N}$

- Dot notation is shorthand for sum over certain values in a matrix M
- $M_{i.} = \sum_j^c M_{ij}$
- $M_{.j} = \sum_i^r M_{ij}$

Degrees of freedom (df): $(r - 1) \times (c - 1)$

Grieve 2007: Quantitative Authorship Attribution: An Evaluation of Techniques

Zarah Weiß

Introduction

Textual Measurements

Length Measures

Vocabulary Richness Measures

Frequency Measures

The Algorithm

The Corpus

Experiment & Results

Experiment

Results

Combination of Techniques

Conclusion

References

Discussion

The Algorithm

Statistics

- H_0 assumes independence
- Two-sided, non-directional test
- Lower chi-square score indicates similarity
- If 0, identical sets
- Else: Consult critical chi-square table (not in Grieve 2007)

The Corpus

Prerequisites

Goal: compile a representative corpus

- Representativeness **not** in terms of variety of an author's language
- Representativeness in terms of the anonymous text
- Representativeness in terms of **idiolects** of the respective authors

Idiolect:

- Often used as "variety of language that encompasses the totality of an individual's utterances" (Grieve 2007:255)
- Originally: "totality of the possible utterances of one speaker at one time in using a language to interact with one other speaker" (Hockett 1948:7)

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

The Corpus

Realisation

The corpus:

- Samples from London *Telegraph*'s opinion columns
- Freely available in online archive
- 40 authors with 40 columns each !
- Comparable and challenging text length: 500 to 2,000 words
- Mostly time span from Jan. 2004 to Jan. 2005 (all from 2000 to 2005)
- Different subjects due to same time span

Controlled for:

- Within authors: Register, audience, production time, dialect
- Across authors: See above, also: age, social background

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Experiment & Results

Experiment

Test for each textual measure:

1. Select an author
2. Select a text by this author $\rightarrow$ anonymous text
3. Run attribution algorithm
4. Continue until all texts by all authors have been attributed
5. Calculate success rate of textual measure: $\frac{\#\ successful\ \ attributions}{\#\ attempted\ \ attributions}$

Experiment & Results

Experiment

Varying tests:

- Each textual measure tested for 40, 20, 10, 5, 4, 3, and 2 possible authors
- Each test with less than 40 possible authors repeated 200 times with random samples from set of possible authors
- Same 200 random samples for N possible authors used for each measure
- For repeated tests success rates were averaged

Evaluation:

- Relative accuracy
- Successful if at least 75% accuracy

Experiment & Results

Word- and Sentence-Length

Table 2 Word- and sentence-length results

Textual measurement			Test accuracy (%)						
Type	Variant		Possible authors						
	Unit	Range	40	20	10	5	4	3	2
Average word-length	Grapheme		7	12	22	39	46	55	70
Average sentence-length	Word		6	11	21	37	44	53	69
Average sentence-length	Grapheme		6	12	22	39	45	53	70
Word-length profile	one grapheme	1–15 characters	18	26	39	54	60	68	79
Word-length profile	one grapheme	1–5 characters	11	18	29	45	51	60	74
Sentence-length profile	five words	1–50 words	11	18	29	44	51	60	74
Sentence-length profile	five words	1–30 words	8	16	26	41	47	57	71
Sentence-length profile	ten words	1–50 words	10	17	28	44	50	59	73
Sentence-length profile	ten words	1–30 words	8	14	24	38	45	54	70
Sentence-length profile	twenty-five characters	1–300 characters	12	20	31	46	53	62	74
Sentence-length profile	twenty-five characters	1–200 characters	10	17	28	43	50	59	73
Sentence-length profile	fifty characters	1–300 characters	11	19	30	45	52	61	74
Sentence-length profile	fifty characters	1–200 characters	9	16	26	41	48	57	72

Figure: Grieve 2007:259.

Experiment & Results

Vocabulary Richness

Table 3 Vocabulary richness results

Textual measurement	Test accuracy (%)						
	Possible authors						
	40	20	10	5	4	3	2
Unrestricted Type–Token ratio	8	16	27	44	51	61	75
Restricted Type–Token ratio	3	7	14	27	33	42	59
Yule's *K* and Simpson's *D*	6	10	18	33	38	49	65
Guiraud's *R*	7	13	24	41	48	58	73
Herdan's *C*	7	14	25	42	49	59	73
Dugast's *k*	8	14	24	41	48	56	72
Honoré's *H*	7	13	23	38	45	54	70
Sichel's *S* and Michéa's *M*	4	9	16	29	35	45	61
Entropy	8	14	24	40	47	56	72
Tuldava's *LN*	11	18	31	49	55	64	77
W (a = −0.165)	11	17	26	40	46	53	68
W (a = −0.172)	11	17	26	40	45	52	67

Figure: Grieve 2007:260.

Experiment & Results

Grapheme Frequency

Table 4 Grapheme frequency results

Textual measurement		Test accuracy (%) Possible authors						
Type	**Variant**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
Grapheme profile		25	35	47	62	67	74	83
Single-position grapheme profile	1st grapheme in word	20	30	41	56	62	69	80
Single-position grapheme profile	2nd grapheme in word	20	29	41	56	62	69	80
Single-position grapheme profile	3rd grapheme in word	16	24	35	49	55	63	75
Single-position grapheme profile	Last grapheme in word	27	36	49	63	68	73	84
Single-position grapheme profile	2nd to last graph in word	23	31	43	57	63	70	81
Single-position grapheme profile	3rd to last graph in word	19	28	41	56	61	69	80
Multiposition grapheme profile	1st three graphemes in word	34	44	56	69	73	79	87
Multiposition grapheme profile	1st six graphemes in word	43	53	64	76	79	84	90
Multiposition grapheme profile	Last three graphs in word	31	41	53	67	72	77	86
Multiposition grapheme profile	Last six graphs in word	42	52	63	74	79	83	90
Multiposition grapheme profile	First and last six graphs	49	58	68	79	82	86	92
Word-internal grapheme profile		28	39	51	65	70	76	85

Figure: Grieve 2007:260.

Experiment & Results

Word Frequency

Table 5 Word frequency results

Textual measurement		Test accuracy (%)						
		Possible authors						
Type	**Limit**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
Word profile	In at least two texts per author	44	53	63	73	77	82	88
Word profile	In at least five texts per author	48	57	67	77	80	85	88
Word profile	In at least ten texts per author	45	54	64	75	79	84	90
Word profile	In at least fifteen texts per author	40	50	61	73	77	81	88
Word profile	In at least twenty texts per author	39	48	59	71	75	80	88
Word profile	In at least twenty-five texts per author	36	46	58	70	74	80	87
Word profile	In at least thirty texts per author	33	44	56	70	74	79	87
Word profile	In at least forty texts per author	16	23	35	50	57	64	57

Figure: Grieve 2007:261.

Experiment & Results

Positional Stylometry

Table 7 Positional stylometry results

Textual measurement		Test accuracy (%)						
		Possible authors						
Type	**Variant**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
Single-position word profile	1st word in sentence	17	30	36	50	56	64	75
Single-position word profile	2nd word in sentence	11	18	27	41	47	56	69
Single-position word profile	3rd word in sentence	7	13	21	35	41	50	64
Single-position word profile	4th word in sentence	6	10	17	30	35	45	59
Single-position word profile	Last word in sentence	4	7	13	25	30	39	56
Single-position word profile	2nd to last word in sentence	6	11	18	31	37	46	61
Single-position word profile	3rd to last word in sentence	6	10	17	29	35	43	59
Single-position word profile	4th to last word in sentence	7	11	19	31	36	45	60
Multi-position word profile	First four words in sentence	22	31	41	55	60	67	77
Multi-position word profile	First eight words in sentence	19	27	38	51	57	63	75
Multi-position word profile	Last four words in sentence	10	15	24	37	43	51	65
Multi-position word profile	Last eight words in sentence	11	16	25	38	43	52	65
Collocation profile	two words	17	24	34	48	54	61	74
Collocation profile	three words	3	6	11	21	27	35	53

Figure: Grieve 2007:263.

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Experiment & Results

Punctuation Mark Frequency

Table 6 Punctuation mark frequency results

Textual measurement		Test accuracy (%) Possible authors						
Type	**Variant/limit**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
Punctuation mark profile	By punctuation marks	30	40	53	67	71	77	86
Punctuation mark profile	By words	34	45	57	71	75	80	88
Punctuation mark profile	By characters	34	46	58	72	76	80	89
Grapheme and punctuation profile		50	60	70	81	84	87	93
Word and punctuation profile	In at least five texts per author	63	72	80	87	89	92	95
Word and punctuation profile	In at least ten texts per author	61	69	77	86	88	91	95
Word and punctuation profile	In at least twenty texts per author	57	66	75	80	83	87	94

Figure: Grieve 2007:262.

Experiment & Results

N-Gram Frequency

Table 8 *N*-gram frequency results

Textual measurement		Test accuracy (%) Possible authors						
Type	**Limit**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
2-gram profile	In at least two texts per author	58	69	77	84	86	89	94
2-gram profile	In at least ten texts per author	65	72	79	86	88	91	94
2-gram profile	In at least twenty texts per author	60	69	77	85	87	90	94
3-gram profile	In at least two texts per author	56	68	75	82	85	89	92
3-gram profile	In at least ten texts per author	61	70	78	85	88	91	94
3-gram profile	In at least twenty texts per author	61	71	77	85	88	91	94
4-gram profile	In at least two texts per author	56	64	72	81	84	88	92
4-gram profile	In at least ten texts per author	55	64	73	83	85	89	93
4-gram profile	In at least twenty texts per author	49	58	68	78	82	86	91
5-gram profile	In at least two texts per author	45	54	66	77	80	84	90
5-gram profile	In at least ten texts per author	47	55	66	76	79	84	90
5-gram profile	In at least twenty texts per author	34	43	54	67	71	78	85
6-gram profile	In at least two texts per author	35	46	57	70	73	78	86
6-gram profile	In at least ten texts per author	35	45	56	68	72	78	86
6-gram profile	In at least twenty texts per author	23	31	42	56	61	68	79
7-gram profile	In at least two texts per author	34	42	45	59	64	69	81
7-gram profile	In at least ten texts per author	19	26	38	52	57	65	75
7-gram profile	In at least twenty texts per author	12	19	29	44	49	58	71
8-gram profile	In at least two texts per author	18	24	36	50	55	62	74
8-gram profile	In at least ten texts per author	9	16	25	40	46	54	68
8-gram profile	In at least twenty texts per author	7	12	21	35	41	49	66
9-gram profile	In at least two texts per author	12	18	28	41	46	55	68
9-gram profile	In at least ten texts per author	6	11	19	32	38	46	62
9-gram profile	In at least twenty texts per author	4	8	15	28	33	42	60

Figure: Grieve 2007:264.

Experiment & Results

Overall Results

Table 9 Overall results

Textual measurement (Variant)	Test accuracy (%)						
	Possible authors						
	40	20	10	5	4	3	2
Word and punctuation mark profile (5-limit)	63	72	80	87	89	92	95
2-gram profile (10-limit)	65	72	79	86	88	91	94
3-gram profile (10-limit)	61	72	78	85	88	91	94
4-gram profile (10-limit)	55	64	73	83	85	89	93
Grapheme and punctuation mark profile	50	60	70	81	84	87	93
Multiposition graph profile (first and last six in word)	49	58	68	79	82	86	92
Word profile (5-limit)	48	57	67	77	80	85	88
5-gram profile (10-limit)	47	55	66	76	79	84	90
Multiposition grapheme profile (first six in word)	43	53	64	76	79	84	90
Multiposition grapheme profile (last six in word)	42	52	63	74	79	83	90
Punctuation mark profile (by character)	34	46	58	72	76	80	89
6-gram profile (10-limit)	35	45	56	68	72	78	86
Word-internal grapheme profile	28	39	51	65	70	76	85
Single-position grapheme profile (last in word)	27	36	49	63	68	73	84
Grapheme profile	25	35	47	62	67	74	83
7-gram profile (2-limit)	34	42	45	59	64	69	81
Single-position graph profile (2nd to last in word)	23	31	43	57	63	70	81
Single-position grapheme profile (1st in word)	20	30	41	56	62	69	80
Multiposition word profile (first four in sentence)	22	31	41	55	60	67	77
Word-length profile (fifteen intervals of one character)	18	26	39	54	60	68	79
Single-position word profile (1st word in sentence)	17	30	36	50	56	64	75
8-gram profile (2-limit)	18	24	36	50	55	62	74
2-word collocation profile	17	24	34	48	54	61	74
Tuldava's *LN*	11	18	31	49	55	64	77
Sentence-length profile (twelve intervals of twenty-five characters)	12	20	31	46	53	62	74
Sentence-length profile. (ten intervals of five words)	10	17	28	44	50	59	73
9-gram profile (2-limit)	12	18	28	41	46	55	68
Type–Token ratio	8	16	27	44	51	61	75
Herdan's *C*	7	14	25	42	49	59	73
Guiraud's *R*	7	13	24	41	48	58	73
Average word-length	7	12	22	39	46	55	70
Average sentence-length (in characters)	6	12	22	39	45	53	70
Average sentence-length (in words)	6	11	21	37	44	53	69
Yule's *K* and Simpson's *D*	6	10	18	33	38	49	65

Figure: Grieve 2007:265

Combination of 16 measures

5 best performing measures:

- I.e. punctuation, grapheme, word and n-gram frequencies
- Over 75% for up to 5 authors each

9 measures for broader range:

- Length measure: Word- and sentence length distribution in characters
- Vocabulary richness: Tuldava's LN and TTR
- Grapheme frequencies: word-internal grapheme profile
- Punctuation profile: simple punctuation profile
- Positional stylometry: multi-position word and 2-word collocation profiles

Experiment & Results

Combination of Techniques

Table 10 Combination algorithm results

Textual measurement (Variant)	Test accuracy (%)						
	Possible authors						
	40	20	10	5	4	3	2
Weighted combination	69	78	85	91	93	95	97
Simple combination	58	72	82	90	92	94	96
Word and punctuation mark profile (5-limit)	63	72	80	87	89	92	95
2-gram profile (10-limit)	65	72	79	86	88	91	94

Figure: Grieve 2007:267.

Conclusion

Grieve 2007's Conclusion

General evaluation procedure:

- Find reasonable set of possible authors with respect to anonymous text
- Gather representative data set from those authors with respect to anonymous text
- Test wide range of attribution algorithms to determine the best for data set
- Test various weighted variations of best algorithms
- Then perform authorship attribution

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

References

Grieve, Jack (2007). "Quantitative Authorship Attribution: An Evaluation of Techniques". In: *Literary and Linguistic Computing* 22.3, pp. 251–270.

McCarthy, Philip and Scott Jarvis (2007). "A theoretical and empirical evaluation of vocd." In: *Language Testing* 24, pp. 459–488.

McCarthy, Philip and Scott Jarvis (2010). "Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment". In: *Behavior Research Methods* 42.2, pp. 381–392.

Stamatatos, Efstathios (2009). "A Survey of Modern Authorship Attribution Methods". In: *Journal of the American Society for Information Science and Technology* 60.3, pp. 538–556.

Introduction

Textual Measurements

Length Measures

Vocabulary Richness Measures

Frequency Measures

The Algorithm

The Corpus

Experiment & Results

Experiment

Results

Combination of Techniques

Conclusion

References

Discussion

Thank you for your attention!

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion

Discussion

Discussion Pointers

- Is the definition of "words" used in Grieve 2007 reasonable?
  - "continuous string of graphemes and / or digits"
- Concerning the given results, would it seem promising to measure syllable frequencies, too?
- Is the fixed, "arbitrary" (Grieve 2007:264) 75% accuracy mark reasonable for up to 40 authors (random baseline 2.5%)?
- Can we – based on the results – actually conclude, that "positional stylometry measurements have proven to be poor indicators of authorship." (Grieve 2007:263), although the experiment was restricted to a highly specific corpus (newspaper columns)?
- Why would we use chi-square on single measure profiles, when there are classification algorithms that can deal with features of different scales? Especially for multi-measure models.

Introduction
Textual Measurements
Length Measures
Vocabulary Richness Measures
Frequency Measures
The Algorithm
The Corpus
Experiment & Results
Experiment
Results
Combination of Techniques
Conclusion
References
Discussion