Forensic authorship attribution for small texts

Ol'ga Feiguina
Cherches and Associates

Graeme Hirst
University of Toronto

Supported by the Natural Sciences and Engineering Research Council of Canada

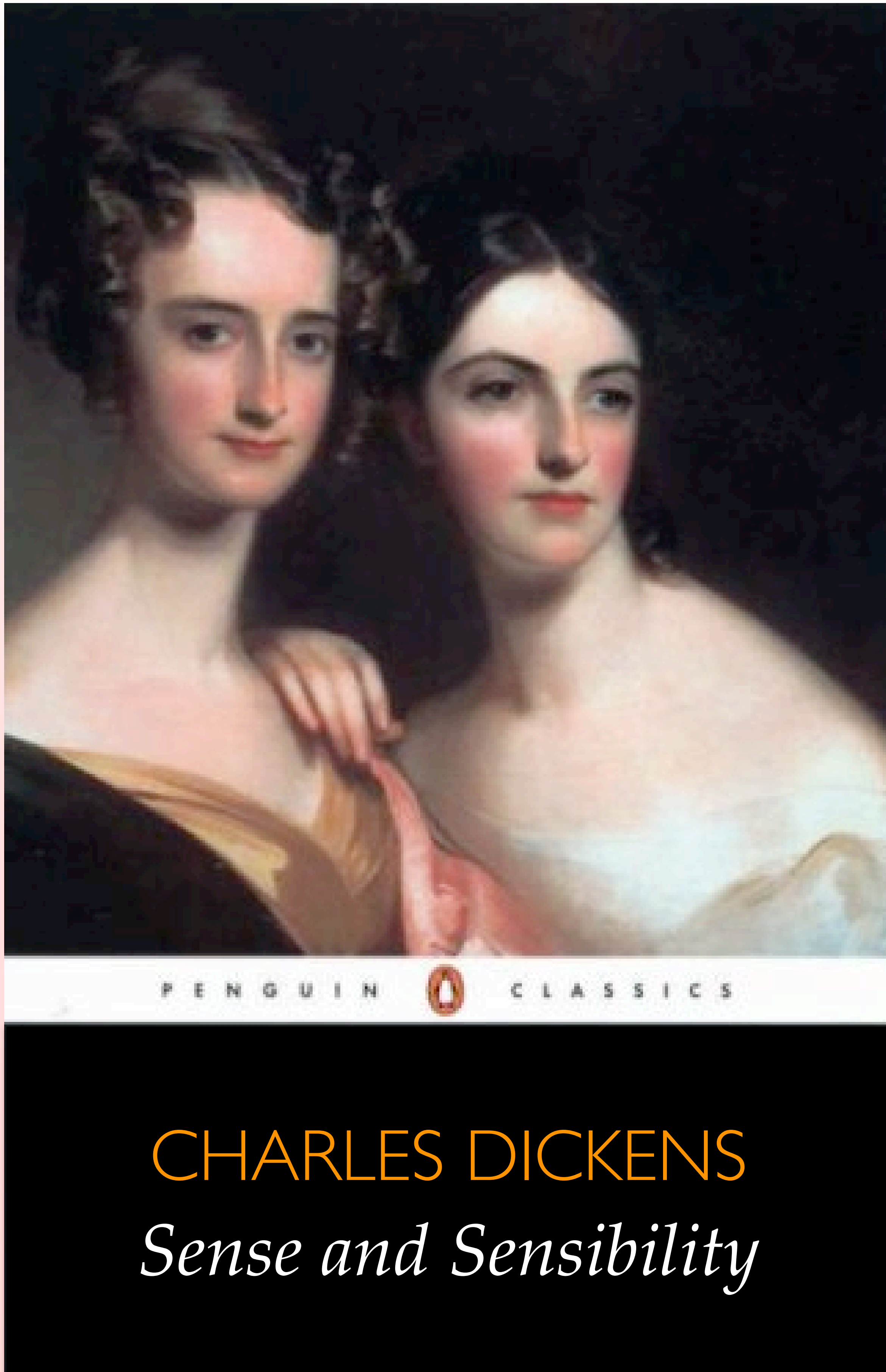
Ol’ga Feiguina

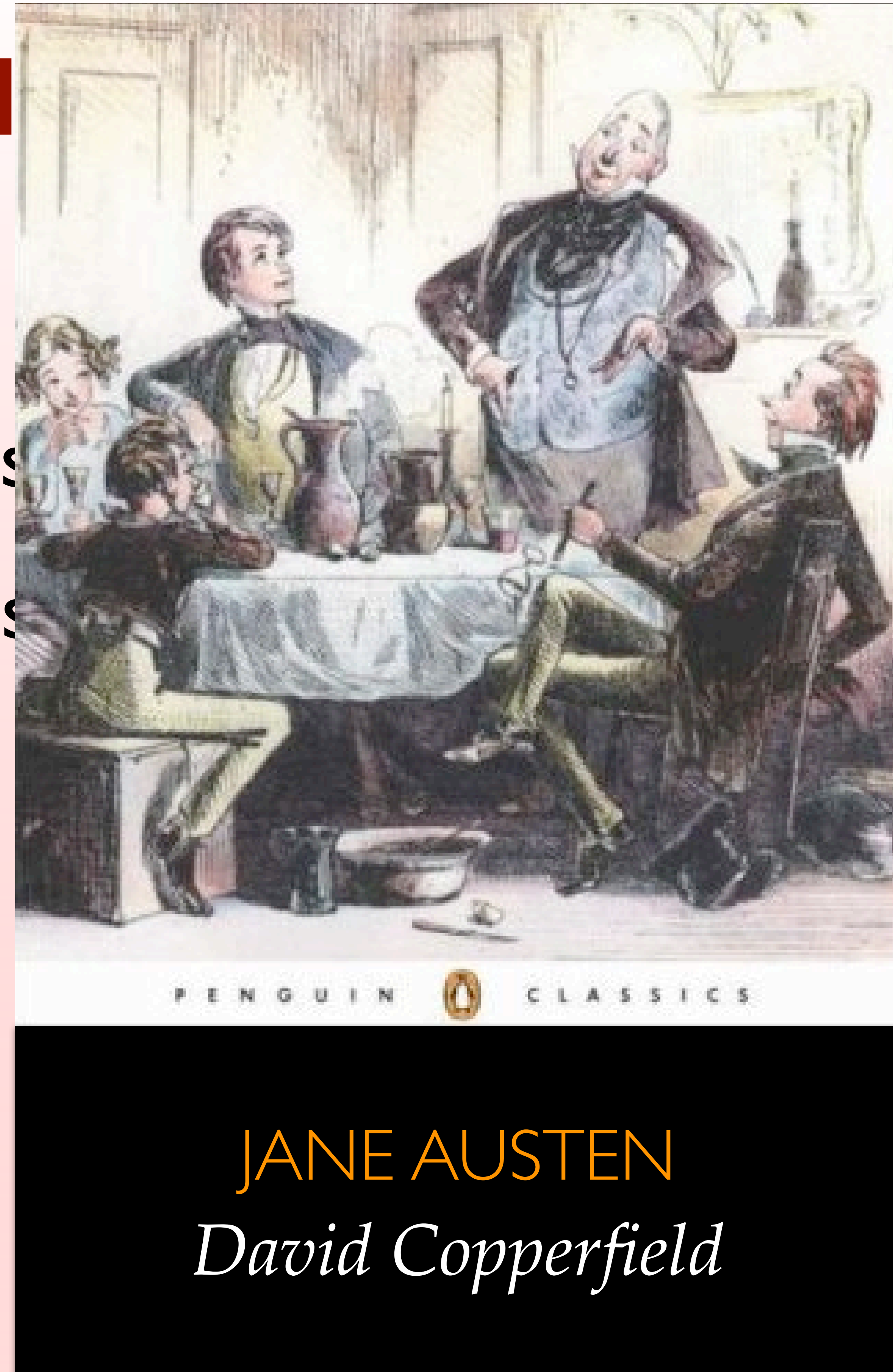
Canonical authorship attribution

- Long literary texts (usually)

Canonical authorship attribution

- Long literary texts (usually)
- Simple methods may suffice


PENGUIN CLASSICS
CHARLES DICKENS
Sense and Sensibility

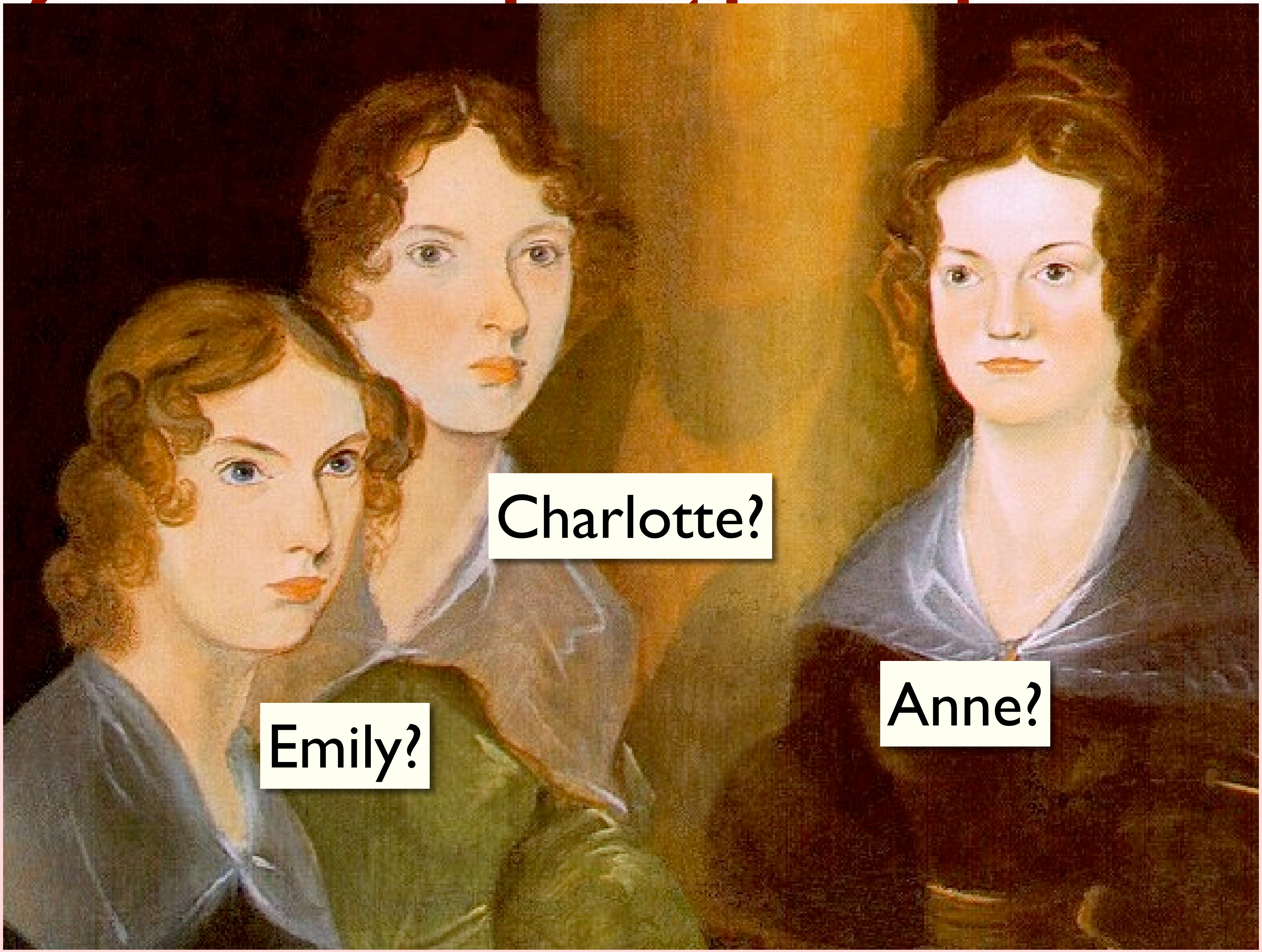

PENGUIN CLASSICS
JANE AUSTEN
David Copperfield

Canonical authorship attribution

- Long literary texts (usually)
- Simple methods may suffice
  - Letter-bigram frequency discriminates Jane Austen from Charles Dickens

Canonical authorship attribution

- Long literary texts (usually)
- Simple methods may suffice
  - Letter-bigram frequency discriminates Jane Austen from Charles Dickens
- Or not


Charlotte?
Emily?
Anne?

Canonical authorship attribution

- Long literary texts (usually)
- Simple methods may suffice
  - Letter-bigram frequency discriminates Jane Austen from Charles Dickens
- Or not
  - Brontë sisters very hard to discriminate (Koppel *et al* 2004)

Koppel, Moshe *et al* (2004). Text categorization for authorship verification. *Eighth International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale.

Short-text authorship attribution

- Literary
- Forensic
- Stylistic consistency checking
    - Writers' aid
    - Forensic

Short-text authorship attribution

- Burrows's Delta: poor results on poems < 500 words
- Zheng *et al*: high accuracy on short domain-specific newsgroup postings

Burrows, John (2002). 'Delta': A measure of stylistic difference and likely authorship. *Literary and Linguistic Computing*, 17(3): 267–287.

Zheng, Rong; Li, Jiexun; Chen, Hsinchun; and Huang, Zan (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3): 378–393.

Short-text authorship discrimination

- Glover and Hirst (1996)
  - Same/diff author judgements
  - Approx 250-word fragments controlled for topic
  - Simple lexical and PoS features
  - Mediocre results

Glover, Angela and Hirst, Graeme (1996). Detecting stylistic inconsistencies in collaborative writing. In: Sharples, Mike and van der Geest, Thea (eds.), *The New Writing Environment*. London: Springer-Verlag, 147–168.

Short-text authorship discrimination

- Graham, Hirst, and Marthi (2005):
  - Neural nets for same/diff author judgements of paragraphs (avg 50 words)
  - PoS tags, lexical features, vocabulary richness
  - Mediocre results

Graham, Neil; Hirst, Graeme; and Marthi, Bhaskara (2005). Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4): 397–415.

The central problem of short texts

The central problem of short texts

They are short

The central problem of short texts

They are short

- Need to use all available information

The central problem of short texts

$\boxed{\text{They are short}}$

- Need to use all available information
- Make better use of syntax, not just PoS

Syntactic structure for authorship attribution

- Baayen *et al*: Sentence as bag of syntactic rewrite rules; vocabulary-richness methods on rules
  - Results better (on long texts) than same method on lexical vocabulary
  - But requires very accurate parsing
  - Applied only to long texts

Baayen, R. Harald; van Halteren, Hans; and Tweedie, Fiona J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3): 121–131.

Syntactic structure for authorship attribution

- Stamatatos *et al*: Chunked texts

Stamatatos, Efstathios; Fakotakis, Nikos; and Kokkinakis, George (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35: 193–214.

Mr. Heathcliff and I are such a suitable pair to divide the desolation between us .

```
NP[Mr. Heathcliff and I] VP[are such]
NP[a suitable pair] VP[to divide]
NP[the desolation] PP[between us] .
```

Syntactic structure for authorship attribution

- Stamatatos *et al*: Chunked texts
- Quantitative features; artefacts of chunker
- Shortish texts (avg 1100 words, half < 1000)
- 10-class accuracy 81%; adding lexical features gives 87%
- Most errors in short texts

Stamatatos, Efstathios; Fakotakis, Nikos; and Kokkinakis, George (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35: 193–214.

Compromise method for small texts

- Robust partial parsing
- Bigrams of syntactic labels as a new feature
- Strengths of both previous methods

Hirst, Graeme and Feiguina, Ol'ga (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, to appear.

Experiments with two Brontë sisters

- Charlotte *vs* Anne: 250,000 words each
- Texts of 1000, 500, or 200 words
  (plus remainder of sentence)
- Support-vector machines; 10-fold cross-validation

Partial parsing

- More than single-level chunking, less than complete syntactic structure (Abney 1996)

Abney, Steven (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4): 337–344.

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

```
[vp [vx [vb Let]]]
[c [c0 [nx [prp it]] [vx [be be]]] [nx [prp theirs]]]
[infp [inf [to to] [vb conceive]]
  [ng [nx [dt the] [nn delight]] [of of] [nx [nn joy]]]]
[vnp [vnx [vbn born]]
  [ax [rb again] [jj fresh]]
  [in out]
  [pp [of of] [nx [jj great] [nn terror]]]]
[cma ,]
[ng [nx [dt the] [nn rapture]] [of of] [nx [nn rescue]]]
[pp [in from] [nx [nn peril]]]
[cma ,]
[nx [dt the] [jj wondrous] [nn reprieve]]
[pp [in from] [nx [nn dread]]]
[cma ,]
[ng [nx [dt the] [nn fruition]] [of of] [nx [nn return]]]
[per .]
```

Partial parsing

- More than single-level chunking, less than complete syntactic structure (Abney 1996)
- Non-recursive, deterministic, fast, robust
- Abney's CASS parser:
    - Cascade of finite-state grammars, one for each level

Abney, Steven (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4): 337–344.

Feature sets

- Syntactic features:
  - Frequencies of bigrams of syntactic labels from CASS

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

```
[vp [vx [vb Let]]]
[c [c0 [nx [prp it]] [vx [be be]]] [nx [prp theirs]]]
[infp [inf [to to] [vb conceive]]
  [ng [nx [dt the] [nn delight]] [of of] [nx [nn joy]]]]
[vnp [vnx [vbn born]]
  [ax [rb again] [jj fresh]]
  [in out]
  [pp [of of] [nx [jj great] [nn terror]]]]
[cma ,]
[ng [nx [dt the] [nn rapture]] [of of] [nx [nn rescue]]]
[pp [in from] [nx [nn peril]]]
[cma ,]
[nx [dt the] [jj wondrous] [nn reprieve]]
[pp [in from] [nx [nn dread]]]
[cma ,]
[ng [nx [dt the] [nn fruition]] [of of] [nx [nn return]]]
[per .]
```

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

```
vp   vx   vb Let
c   c0   nx   prp it     vx   be be       nx   prp theirs
infp   inf   to to    vb conceive
  ng   nx   dt the    nn delight     of of    nx   nn joy
vnp   vnx   vbn born
  ax   rb again    jj fresh
  in out
  pp   of of    nx   jj great    nn terror
cma ,
ng   nx   dt the    nn rapture     of of    nx   nn rescue
pp   in from    nx   nn peril
cma ,
nx   dt the    jj wondrous    nn reprieve
pp   in from    nx   nn dread
cma ,
ng   nx   dt the    nn fruition     of of    nx   nn return
per .
```

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

```
vp  vx  vb
c  c0  nx  prp        vx  be          nx  prp
infp  inf  to      vb
  ng  nx  dt        nn                of        nx  nn
vnp  vnx  vbn
  ax  rb        jj
  in
  pp  of      nx  jj          nn
cma
ng  nx  dt        nn                of        nx  nn
pp  in        nx  nn
cma
nx  dt      jj              nn
pp  in        nx  nn
cma
ng  nx  dt        nn                of        nx  nn
per
```

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

```
vp vx vb
c c0 nx prp vx be nx prp
infp inf to vb
ng nx dt nn of nx nn
vnp vnx vbn
ax rb jj
in
pp of nx jj nn
cma
ng nx dt nn of nx nn
pp in nx nn
cma
nx dt jj nn
pp in nx nn
cma
ng nx dt nn of nx nn
per
```

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

```
vp vx vb c c0 nx prp vx be nx prp infp inf to vb ng nx dt
nn of nx nn vnp vnx vb n ax rb jj in pp of nx jj nn cma ng
nx dt nn of nx nn pp in nx nn cma nx dt jj nn pp in nx nn
cma ng nx dt nn of nx nn per
```

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

```
vp vx vb c c0 nx prp vx be nx prp infp inf to vb ng nx dt
nn of nx nn vnp vnx vb n ax rb jj in pp of nx jj nn cma ng
nx dt nn of nx nn pp in nx nn cma nx dt jj nn pp in nx nn
cma ng nx dt nn of nx nn per
```

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

vp $\boxed{\texttt{vx vb}}$ c c0 nx prp vx be nx prp infp inf to vb ng nx dt nn of nx nn vnp vnx vb n ax rb jj in pp of nx jj nn cma ng nx dt nn of nx nn pp in nx nn cma nx dt jj nn pp in nx nn cma ng nx dt nn of nx nn per

Let it be theirs to conceive the delight of joy born again fresh out of great terror , the rapture of rescue from peril , the wondrous reprieve from dread , the fruition of return .

```
vp vx vb c c0 nx prp vx be nx prp infp inf to vb ng nx dt
nn of nx nn vnp vnx vb n ax rb jj in pp of nx jj nn cma ng
nx dt nn of nx nn pp in nx nn cma nx dt jj nn pp in nx nn
cma ng nx dt nn of nx nn per
```

Feature sets

- Syntactic features:
    - Frequencies of bigrams of syntactic labels from CASS

Feature sets

- Syntactic features:
    - Frequencies of bigrams of syntactic labels from CASS
    - Frequencies of rewrite rules from CASS

Feature sets

- Syntactic features:
  - Frequencies of bigrams of syntactic labels from CASS
  - Frequencies of rewrite rules from CASS
  - Vocabulary richness measures on rewrite rules (à la Baayen *et al*)

Feature sets

- Standard lexical features
  - Frequency of function words, punctuation, *i*-letter words, *i*-syllable words, …
  - Avg word length, sentence length,…
  - Vocabulary richness measures
- In-between features
  - Frequency of PoS tags

Results

Classification accuracy (in percent)*

	Text size		
	1000	500	200
All syntactic features	**99.5**	94.2	87.5
Label bigram freqs	99.0	93.4	84.9
Rule freqs	93.2	93.4	83.8
Vocab-richness on rules	76.6	76.7	70.3
Label bigram and rule freqs	98.4	95.8	87.4
Lexical features	97.5	90.5	85.6
PoS freqs	93.8	93.4	82.7
Lexical features and PoS freqs	98.9	95.0	89.5
All features	99.2	**96.8**	**92.4**

**Average across 10-fold cross-validation*

Results

Classification accuracy (in percent)*

	Text size		
	1000	500	200
All syntactic features	**99.5**	94.2	87.5
Label bigram freqs	99.0	93.4	84.9
Rule freqs	93.2	93.4	83.8
Vocab-richness on rules	76.6	76.7	70.3
Label bigram and rule freqs	98.4	95.8	87.4
Lexical features	97.5	90.5	85.6
PoS freqs	93.8	93.4	82.7
Lexical features and PoS freqs	98.9	95.0	89.5
All features	99.2	**96.8**	**92.4**

**Average across 10-fold cross-validation*

Results

Classification accuracy (in percent)*

	Text size		
	1000	500	200
All syntactic features	**99.5**	94.2	87.5
Label bigram freqs	99.0	93.4	84.9
Rule freqs	93.2	93.4	83.8
Vocab-richness on rules	76.6	76.7	70.3
Label bigram and rule freqs	98.4	95.8	87.4
Lexical features	97.5	90.5	85.6
PoS freqs	93.8	93.4	82.7
Lexical features and PoS freqs	98.9	95.0	89.5
All features	99.2	**96.8**	**92.4**

**Average across 10-fold cross-validation*

Results

Classification accuracy (in percent)*

	Text size		
	1000	500	200
All syntactic features	**99.5**	94.2	87.5
Label bigram freqs	99.0	93.4	84.9
Rule freqs	93.2	93.4	83.8
Vocab-richness on rules	76.6	76.7	70.3
Label bigram and rule freqs	98.4	95.8	87.4
Lexical features	97.5	90.5	85.6
PoS freqs	93.8	93.4	82.7
Lexical features and PoS freqs	98.9	95.0	89.5
All features	99.2	**96.8**	**92.4**

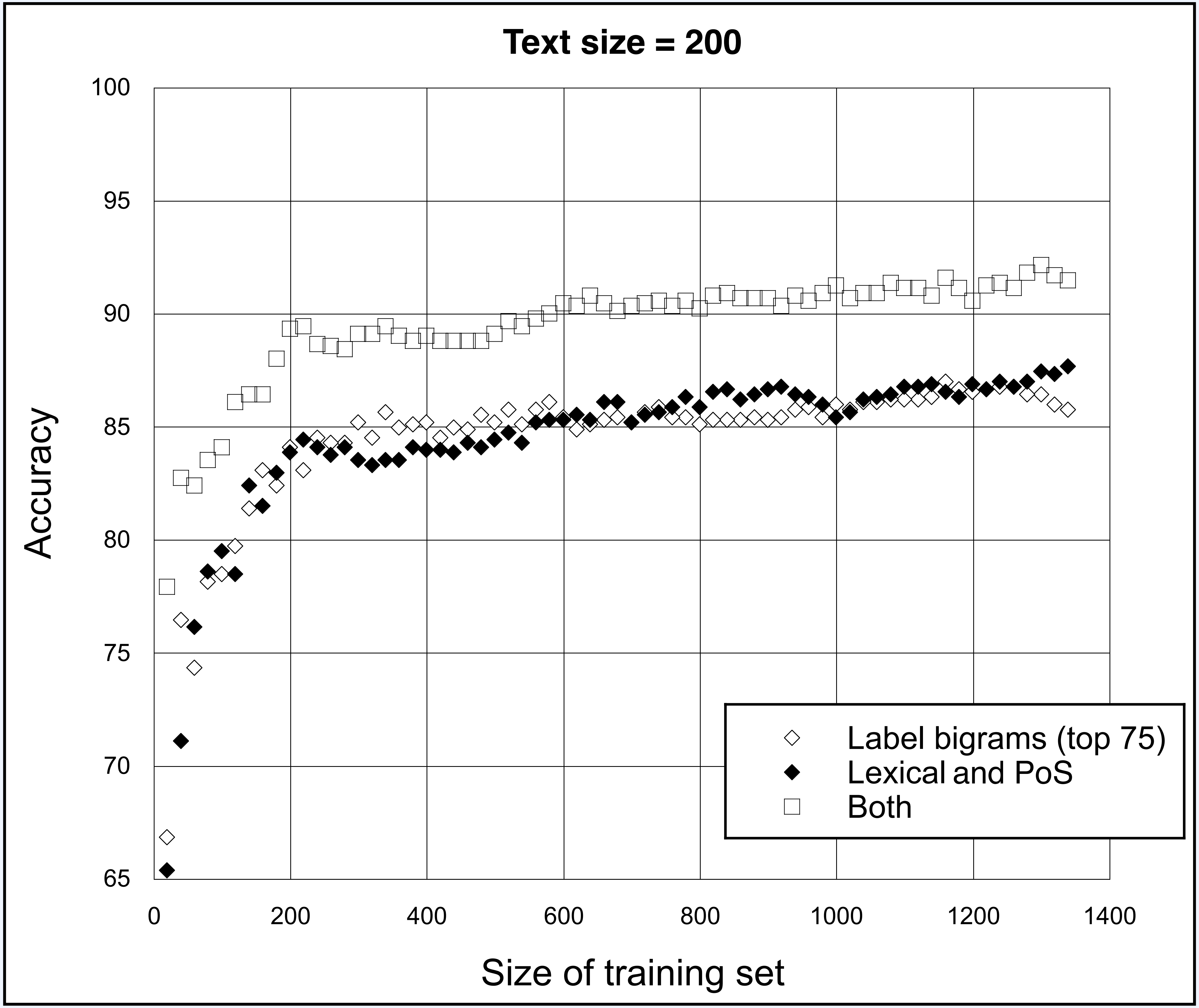
**Average across 10-fold cross-validation*

Results

Classification accuracy (in percent)*

	Text size		
	1000	500	200
All syntactic features	**99.5**	94.2	87.5
Label bigram freqs	99.0	93.4	84.9
Rule freqs	93.2	93.4	83.8
Vocab-richness on rules	76.6	76.7	70.3
Label bigram and rule freqs	98.4	95.8	87.4
Lexical features	97.5	90.5	85.6
PoS freqs	93.8	93.4	82.7
Lexical features and PoS freqs	98.9	95.0	89.5
All features	99.2	**96.8**	**92.4**

**Average across 10-fold cross-validation*



Text size = 200
Accuracy
Size of training set
100
95
90
85
80
75
70
65
0
200
400
600
800
1000
1200
1400
Label bigrams (top 75)
Lexical and PoS
Both

Most-discriminating label bigrams

`cc c`	Coordinating conjunction followed by clause
`prp cma`	Personal pronoun followed by comma
`name nnp`	Name starting with proper noun
`nx nn`	Noun chunk starting with common noun
`cc vp`	Coordinating conjunction followed by verb phrase
`cma c`	Comma followed by clause
`vb nx`	Verb followed by noun chunk
`uh c`	Interjection followed by clause
`dtp nn`	Determiner followed by noun

Forensic authorship attribution

- E.g., anonymous letters

JERALD AND SANDRA TANNER,

I am writing you anonymousely to tip you off to a cover up by the Mormon church and the document discover Mark Hoffmann.

A few days ago Mark showed me the original actual Egyption Papyrus of the round facsimile of the P. of G. P. It is in many pieces and is pasted onto a piece of heavy paper. There are pencil and ink drawings filling in the missing parts. There is another square piece of papyrus pasted on the same piece of paper. Mark told me not to tell anyone about this. He told me it would never be seen again after the church go it. He is keeping a large color photograph.

I am telling you these things because I do not think it should be covered up and I think you can find out more about it. Mark payed over $1,000 from someone in Texas. Please do not tell ANYONE you were tipped off by this letter. Good Luck.

Tip-off letter (180 words) re forged Mormon document. From *Tracking the White Salamander: The Story of Mark Hofmann, Murder and Forged Mormon Documents*, by Jerald Tanner, 1987. *http://www.utlm.org/onlinebooks/trackingcontents.htm*

JERALD AND SANDRA TANNER,

I am writing you anonymousely to tip you off to a cover up by the Mormon church and the document discover Mark Hoffmann.

round

piece

parts

paper

be se

up an

someo

Good

Tip-of

Salam

Jerald

ANONYMOUS LETTER TO MEMBERS
OF THE FACULTY COMMITTEE ON
ACADEMIC FREEDOM AND TENURE
ARIZONA STATE UNIVERSITY

Dear Sir:

It seems appropriate that you should be informed of one of the most recent activities of Morris J. Starsky. Starsky learned of a suicide attempt by one of his close campus co-workers, David Murphy. Feeling that Murphy could no longer be trusted as a member of the campus socialist group, Starsky demanded that Murphy return all literature and other materials belonging to the socialist group. Murphy refused to give Starsky a quantity of socialist literature in his possession until Starsky would pay him a sum slightly in excess of $50 which was owed for telephone calls charged by Starsky to Murphy's telephone. Morris Starsky was indignant at Murphy's independent attitude and at 2:00 A. M. on April 5, 1970 he, accompanied by his wife Pamela and two young male associates, invaded Murphy's apartment and

Anonymous poison-pen letter (347 words plus heading and signature) from FBI campaign to “neutralize” Socialist Worker’s Party candidate, 1968.
http://www.icdc.com/~paulwolf/cointelpro/cointelindex.htm

Experiments with (simulated) forensic data

- Chaski's writing-sample database
  - 11 writers:
    American English, varied ages, similar background
  - Set of 10 topics:
    Threatening letter, apology, complaint, love letter, …
  - ~2000 words or ~100 sentences per author
  - 73 texts: 4 to 10 per author

Thanks to Carole Chaski for allowing us to use her data!

Chaski, Carole E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence,* 4(1).

Dear Mr. Smith,

It has been brought to my attention that you intend to run for the position of school board member. I cannot believe that someone of your character would even consider this. Because of your past complication in the Jane Brown scandal, I cannot stand idly by and allow you to pursue a position on the board of a public school system. I do not believe your character and total lack of morality would lend itself to the education of our town's children.

Therefore, if I do not read of your withdrawal from the election, by next Tuesday, I will be forced to come forward and reveal my knowledge of your wrong doing. By doing this, I will reluctantly bring scorn and shame on your family- but I will do it because I feel our impressionable children should not in any manner be associated with you. Please don't force me to do this. Drop your name from the race!

Sample 080-10 (177 words)

Chaski's method

- Features (manually assisted):
  - Counts of punctuation classified by edge (clause, phrase, morpheme)
  - Counts of syntactically (un-)marked consituents
  - Average word length
- Classifier: Linear discriminant function analysis

Chaski, Carole E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence,* 4(1).

Chaski, Carole E. (2005). Computational stylistics in forensic author identification. *SIGIR Workshop on Stylistic Analysis of Text for Information Access.*

Chaski's results

- Pairwise classification by author:

  2005 *IJDE*: Accuracy = 95%*
  2005 *SIGIR wkshp*: Accuracy = 81.3%‡

  Average across all author pairs using leave-one-out cross-validation.
  **Using SPSS*
  ‡*Manual replication*

Chaski, Carole E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence,* 4(1).

Chaski, Carole E. (2005). Computational stylistics in forensic author identification. *SIGIR Workshop on Stylistic Analysis of Text for Information Access.*

Our results

Pairwise classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	86.1	78.8
Rule freqs	87.3	72.4
Label bigram and rule freqs	88.3	75.4
Lexical features	84.4	83.2
Label bigrams and lexical features	88.3	83.3
PoS freqs	89.2	84.1
PoS freqs and lexical features	**91.2**	**85.6**
PoS, lexical features, label bigrams	89.3	80.0
All features	88.7	75.6

**Average across all author pairs with 10-fold cross-validation*

Our results

Pairwise classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	86.1	78.8
Rule freqs	87.3	72.4
Label bigram and rule freqs	88.3	75.4
Lexical features	84.4	83.2
Label bigrams and lexical features	88.3	83.3
PoS freqs	89.2	84.1
PoS freqs and lexical features	**91.2**	**85.6**
PoS, lexical features, label bigrams	89.3	80.0
All features	88.7	75.6

**Average across all author pairs with 10-fold cross-validation*

Our results

Pairwise classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	86.1	78.8
Rule freqs	87.3	72.4
Label bigram and rule freqs	88.3	75.4
Lexical features	84.4	83.2
Label bigrams and lexical features	88.3	83.3
PoS freqs	89.2	84.1
PoS freqs and lexical features	**91.2**	**85.6**
PoS, lexical features, label bigrams	89.3	80.0
All features	88.7	75.6

**Average across all author pairs with 10-fold cross-validation*

Our results

Pairwise classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	86.1	78.8
Rule freqs	87.3	72.4
Label bigram and rule freqs	88.3	75.4
Lexical features	84.4	83.2
Label bigrams and lexical features	88.3	83.3
PoS freqs	89.2	84.1
PoS freqs and lexical features	**91.2**	**85.6**
PoS, lexical features, label bigrams	89.3	80.0
All features	88.7	75.6

**Average across all author pairs with 10-fold cross-validation*

Our results

Pairwise classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	86.1	78.8
Rule freqs	87.3	72.4
Label bigram and rule freqs	88.3	75.4
Lexical features	84.4	83.2
Label bigrams and lexical features	88.3	83.3
PoS freqs	89.2	84.1
PoS freqs and lexical features	**91.2**	**85.6**
PoS, lexical features, label bigrams	89.3	80.0
All features	88.7	75.6

Average across all author pairs with 10-fold cross-validation

Most-discriminating label bigrams

47	`vb vp`	Verb followed by verb phrase
39	`in dt-a`	Preposition followed by determiner *a*
38	`jjr nn`	Comparative adjective followed by noun
35	`hvd rb`	*had* followed by adverb
35	`dtp-q nns`	{*all*, *some*} followed by plural noun
33	`vnx rb`	Past tense verb group starting with adverb
27	`ber vbg`	*are* followed by progressive verb
24	`ben nx`	*been* followed by noun chunk
23	`bedr vbg`	*were* followed by progressive verb
20	`tunit nx`	Time-unit word followed by noun chunk

↑ *Number of author pairs (out of 55) for which this bigram is in top 10 discriminators*

Our results

Multiclass classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	57.5	39.6
Rule freqs	56.2	25.5
Label bigram and rule freqs	56.2	34.9
Lexical features	41.4	**50.9**
Label bigrams and lexical features	**60.3**	48.1
PoS freqs	**60.3**	34.0
PoS freqs and lexical features	51.0	49.1
PoS, lex features, label bigrams	58.9	50.0
All features	**60.3**	37.7

**Average across all authors with 10-fold cross-validation*

Our results

Multiclass classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	57.5	39.6
Rule freqs	56.2	25.5
Label bigram and rule freqs	56.2	34.9
Lexical features	41.4	**50.9**
Label bigrams and lexical features	**60.3**	48.1
PoS freqs	**60.3**	34.0
PoS freqs and lexical features	51.0	49.1
PoS, lex features, label bigrams	58.9	50.0
All features	**60.3**	37.7

**Average across all authors with 10-fold cross-validation*

Our results

Multiclass classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	57.5	39.6
Rule freqs	56.2	25.5
Label bigram and rule freqs	56.2	34.9
Lexical features	41.4	**50.9**
Label bigrams and lexical features	**60.3**	48.1
PoS freqs	**60.3**	34.0
PoS freqs and lexical features	51.0	49.1
PoS, lex features, label bigrams	58.9	50.0
All features	**60.3**	37.7

**Average across all authors with 10-fold cross-validation*

Our results

Multiclass classification accuracy (in percent)*

	Whole docs	200-wd texts
Label bigram freqs	57.5	39.6
Rule freqs	56.2	25.5
Label bigram and rule freqs	56.2	34.9
Lexical features	41.4	**50.9**
Label bigrams and lexical features	**60.3**	48.1
PoS freqs	**60.3**	34.0
PoS freqs and lexical features	51.0	49.1
PoS, lex features, label bigrams	58.9	50.0
All features	**60.3**	37.7

**Average across all authors with 10-fold cross-validation*

Discussion

- Features from partial parsing give accuracy between Chaski's two reported results

Discussion

- Features from partial parsing give accuracy between Chaski's two reported results
- Better results on simulated forensic data than on Brontës at same training set size

Discussion

- Features from partial parsing give accuracy between Chaski's two reported results
- Better results on simulated forensic data than on Brontës at same training set size
  - Most ordinary writers are an easier problem than Brontës?

Discussion

- Features from partial parsing give accuracy between Chaski's two reported results
- Better results on simulated forensic data than on Brontës at same training set size
  - Most ordinary writers are an easier problem than Brontës?
  - Robust partial parsing useful for poorly written texts

Discussion

- Features from partial parsing give accuracy between Chaski's two reported results
- Better results on simulated forensic data than on Brontës at same training set size
  - Most ordinary writers are an easier problem than Brontës?
  - Robust partial parsing useful for poorly written texts
  - But punctuation, sentence-splitting still a problem

Discussion

- But PoS and lexical features perform even better (contra results on Brontës).

Discussion

- But PoS and lexical features perform even better (contra results on Brontës).
  - *Observation:* Most-discriminating label bigrams are more lexical, less "constituent-oriented" on Chaski's data than on Brontës.

Discussion

- But PoS and lexical features perform even better (contra results on Brontës).
  - *Observation:* Most-discriminating label bigrams are more lexical, less "constituent-oriented" on Chaski's data than on Brontës.
- PoS bigrams as feature for future study

Conclusion

Conclusion

- Short texts are short.
  Small datasets are small.

Conclusion

- Short texts are short.
  Small datasets are small.
- Don’t assume methods will generalize across genres or text types