Literary and Linguistic Computing: Motivation and a Prodigious Case Study

W.J. Scheirer
Department of Computer Science
University of Colorado at Colorado Springs

The Part About the Critics…

Warnings

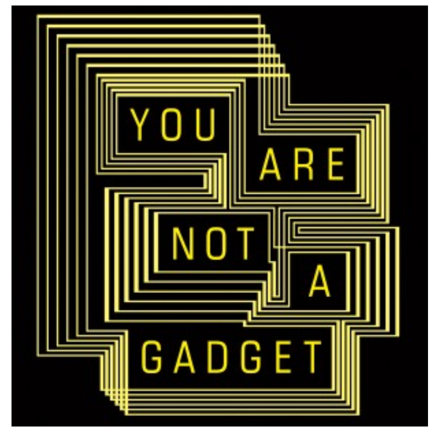
"Regenerations, reproductions, returns, hydras, and medusas do not get us any further… This is evident in current problems in information science and computer science, which still cling to the oldest modes of thought in that they grant all power to a memory or central organ."

Deleuze and Guattari, *A Thousand Plateaus*, Introduction: Rhizome

Warnings

“People degrade themselves all the time in order to make machines seem smart.”

“…a new philosophy: that the computer can understand people better than people can understand themselves.”



“We have repeatedly demonstrated our species’s bottomless ability to lower our standards to make information technology good, but every manifestation of intelligence in a machine is ambiguous.”

Jaron Lanier, “The Serfdom of Crowds,” Harper’s, Feb. 2010

Warnings

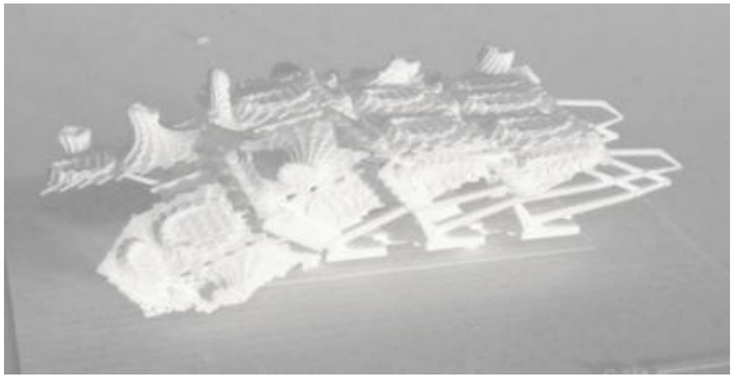
In the early 1960s, it was "envisioned that building a thinking machine would take about a decade."

"By the mid-1980s, many scientists both inside and outside of the artificial intelligence community had come to see the effort as a failure."

NY Times, "Optimism as Artificial Intelligence Pioneers Reunite," Dec. 7, 2009

Inklings

New logics are always still about "questions of logic and existence"



"mathematics and the formalization of discourse"

"information theory and its application to the analysis of life"

Foucault, *The Archaeology of Knowledge*

Inklings

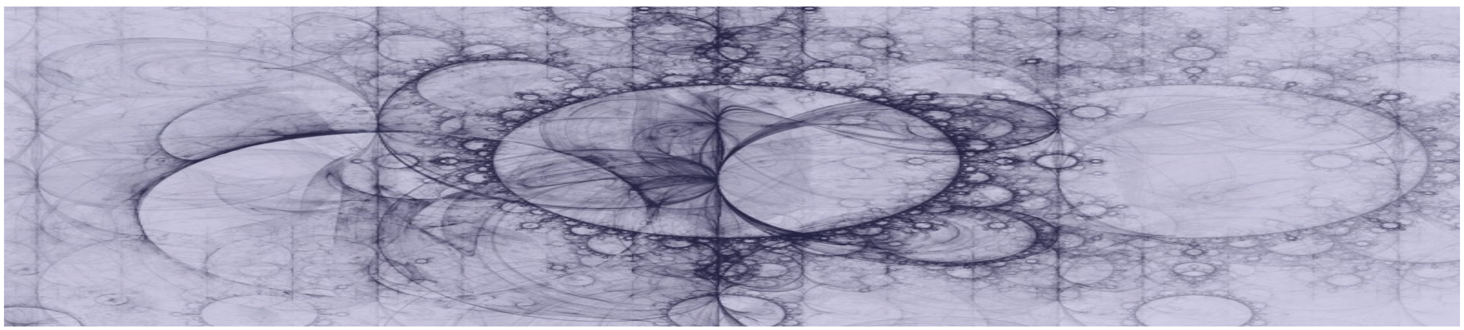
"INFORMATION = ENTROPY"

"Here we have not spoken of information except in the social register of communication. But it would be enthralling to consider this hypothesis even within the parameters of cybernetic information theory."

Jean Baudrillard, *Simulacra and Simulation*, VII. The Implosion of Meaning in the Media

And More Warnings

"And more than one English graduate student has written papers trying to apply information theory to literature -- the kind of phenomenon that later caused Dr. Shannon to complain of what he called a 'bandwagon effect'."



"Information theory has perhaps ballooned to an importance beyond its actual accomplishments."

NY Times, "Claude Shannon, Mathematician, Dies at 84," Feb. 27, 2001

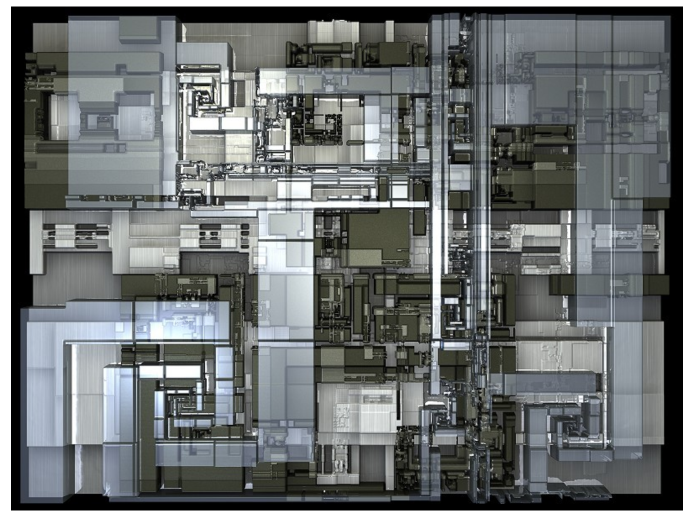
Software Tools

Write programs that do one thing and do it well.

*Especially what you might already be doing by hand.

Software Tools

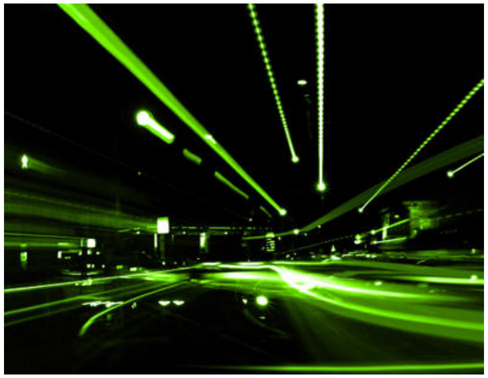
- What types of interesting problems can computers solve?
  - Iteration, Recursion, and Feedback
    - Repetitive loops
  - Collection, Multiplicity, and Parallelism
    - Efficient processing
  - Adaptation, Learning, and Evolution
    - Pattern recognition



Software Tools

- Useful trends in computational linguistics:
  - **Probabilistic Models**
  - **Machine Learning**

Digital Humanities



- Integrate technology into scholarly activity (in a non-gratuitous fashion)
- “knowledge-making, dispersal, and collection”
- Fun interdisciplinary collaboration!

Academic Forums

- Conferences
  - Digital Humanities
    - 2010 Meeting: http://dh2010.cch.kcl.ac.uk/
  - Chicago Colloquium on Digital Humanities and Computer Science
    - 2009 Meeting: http://dhcs.iit.edu/
- Journal
  - Literary and Linguistic Computing: http://llc.oxfordjournals.org/
- Societies
  - The Association for Literary and Linguistic Computing: http://www.allc.org/
  - The Association for Computers in the Humanities: http://www.ach.org/
  - The Society for Digital Humanities: http://www.sdh-semi.org/

A Prodigious Case Study

A Prodigious Case Study

- Forstall and Scheirer 2009[1]
  - “Features From Frequency: Authorship and Stylistic Analysis Using Repetitive Sound”
- A foray into stylistics for literary study
  - Large survey of English, Latin and Greek literature using a common stylistic “tool”.

1. Proc. of the 2009 Chicago Colloquium on Digital Humanities and Computer Science (forthcoming)

Inspiration…

"…He's got *go,* anyhow."

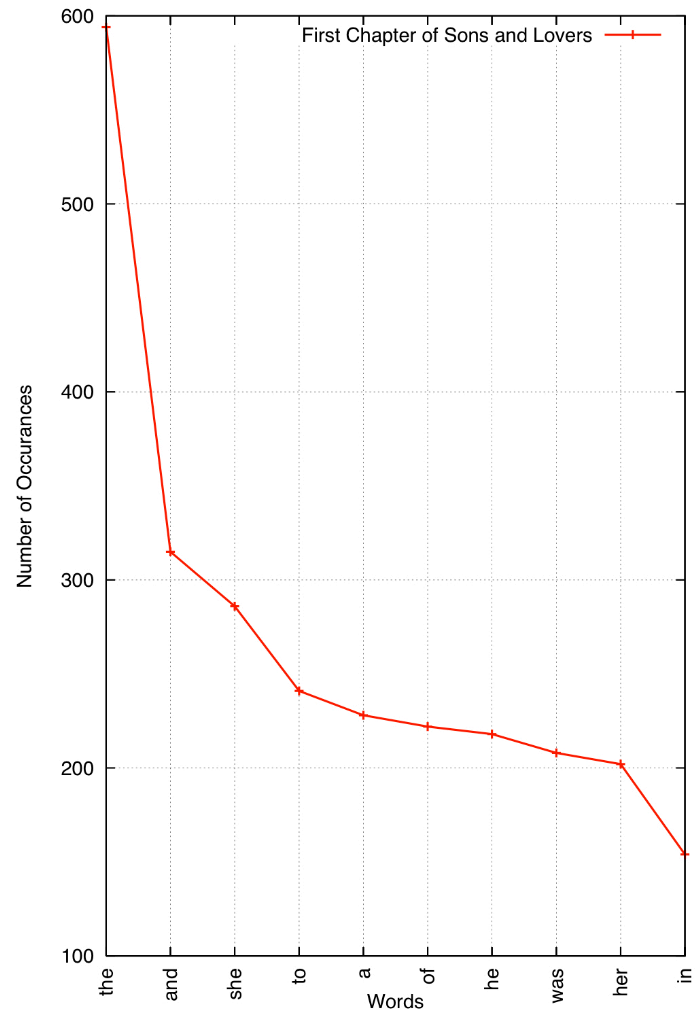
"Certainly, he's got go," said Gudrun. "In fact I've never seen a man that showed signs of so much. The unfortunate thing is, where does his *go* go to, what becomes of it?"

"Oh I know," said Ursula. "It goes in applying the latest appliances!"

Lawrence, *Women in Love,* Chpt. 4

Style Markers

- Function words
  - Zipf's law*:
    - "…in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table"
  - The most frequently used words tend to be articles, adverbs, conjunctions, and pronouns
    - In practice, half of the words in a text occur just once (*hapax legomena*)



*G. Zipf, "Human Behavior and the Principle of Least-Effort," 1949

Style Markers

- n-grams
  - Character-level n-grams capture sound and word information; Phoneme-level n-grams capture pure sound information
  - *Character-level and Phoneme-level* n-grams behave the same way as Word-level n-grams:

$$P(\mathrm{h} \mid \mathrm{t}) = C(\mathrm{th}) \,/\, C(\mathrm{t})$$

  - Generalizing:

$$P(e_n \mid e_{n\text{-}N+1}) = \frac{C(e_{n\text{-}N+1}^{n\text{-}1} e_n)}{C(e_{n\text{-}N+1}^{n\text{-}1})}$$

Functional n-gram

- We need a style marker to capture sound frequency
- Solution:
  - Recall the Zipfian distribution…
    - The n-grams of a text are ranked by frequency, but the features themselves remain the relative n-gram probabilities
- Functional n-grams relieve any need for feature vector normalization
- Functional n-grams are used as direct input for any supervised learning algorithm
  - In this work, we'll use SVM[1] and PCA[2]

1. J. Diederich, J. Kindermann, E. Leopold and G. Paass, "Authorship attribution with Support Vector Machines," *Applied Intelligence*, 19(1-2), pp. 109–123, 2003.

2. D. Holmes, M. Robertson, and R. Paez, "Stephen Crane and the New York Tribune: A Case Study in Traditional and Non-traditional Authorship Attribution," *Computers and the Humanities*, 35(3), pp. 315-331, 2001

Experiments: Authorship Attribution

- The experimental corpus
  - Novels
    - 2 English Novelists
  - Poetry
    - 11 Poets
    - 3 different periods represented
      - Romantic, Renaissance, and Classical
    - Overall, the amount of text is less per poet over a span of works than for a novelist's single long novel.
- 10-fold cross validation
  - Texts for each author split into n sub-samples, and randomly sampled

Experiments: The English Novel

- The English novel corpus

- Austen - *Sense and Sensibility*, 14,731 lines, 118,542 words
- Lawrence - *Sons and Lovers*, 21,978 lines, 160,035 words
- Lawrence - *Women in Love*, 23,029 lines, 176,391 words

Experiments: The English Novel

Test	**Function Words** Training Vectors	**Function Words** % Misclassified	**Functional Char.-level Bi-grams** Training Vectors	**Functional Char.-level Bi-grams** % Misclassified	**Functional Char.-level Tri-grams** Training Vectors	**Functional Char.-level Tri-grams** % Misclassified
Lawrence vs. Austen	90	0.0	100	0.0575	100	0.0275

Test	**Function Words** Training Vectors	**Function Words** % Misclassified
Lawrence vs. Lawrence	100	0.2125

All features have a vector length of 10

Experiments: Poetry

- The poetry corpus

- Byron - Romantic British poet, 18,074 lines, 125,623 words
- Shelley - Romantic British poet, 18,652 lines, 126,383 words
- Coleridge - Romantic British poet, 2,745 lines, 17,614 words
- Keats - Romantic British poet, 2,652 lines, 19,031 words
- Longfellow - Romantic American poet, 6,081 lines, 31,065 words
- Poe - Romantic American poet, 3,082 lines, 17,495 words
- Chapman - Renaissance British poet, 8,872 lines, 71,253 words
- Milton - Renaissance British poet, 10,608 lines, 79,720 words
- Shakespeare - Renaissance British poet and 2,309 lines, 17,489 words
- Ovid - Classical Latin poet, 11,998 lines, 80,328 words
- Vergil - Classical Latin poet, 10,260 lines, 65,686 words

Experiments: English Poetry, *The Challenge*

You gentlemen, by dint of long seclusion
From better company, have kept your own
At Keswick, and through still continued fusion
Of one another's minds at last have grown
To deem, as a most logical conclusion,
That poesy has wreaths for you alone.
There is a narrowness in such a notion,
Which makes me wish you'd change your lakes for ocean.

Byron, Don Juan 37-44

Now Time his dusky pennons o'er the scene
Closes in steadfast darkness, and the past
Fades from our charmed sight. My task is done:
Thy lore is learned. Earth's wonders are thine own,
With all the fear and all the hope they bring.
My spells are past: the present now recurs.
Ah me! a pathless wilderness remains
Yet unsubdued by man's reclaiming hand.

Shelley, Queen Mab 138-145

Experiments: English Poetry, *The Challenge*

- Sample of functional phoneme and character-level bi-grams for Byron and Shelley

0.2694040669200 ah0 n 0.2634725496800
0.4419285274183 dh ah0 0.4683208701563
0.6186898642414 ao1 r 0.5843537414965
0.1369433323703 t uw1 0.1079038768422
0.2185688405797 eh1 n 0.2256212256212

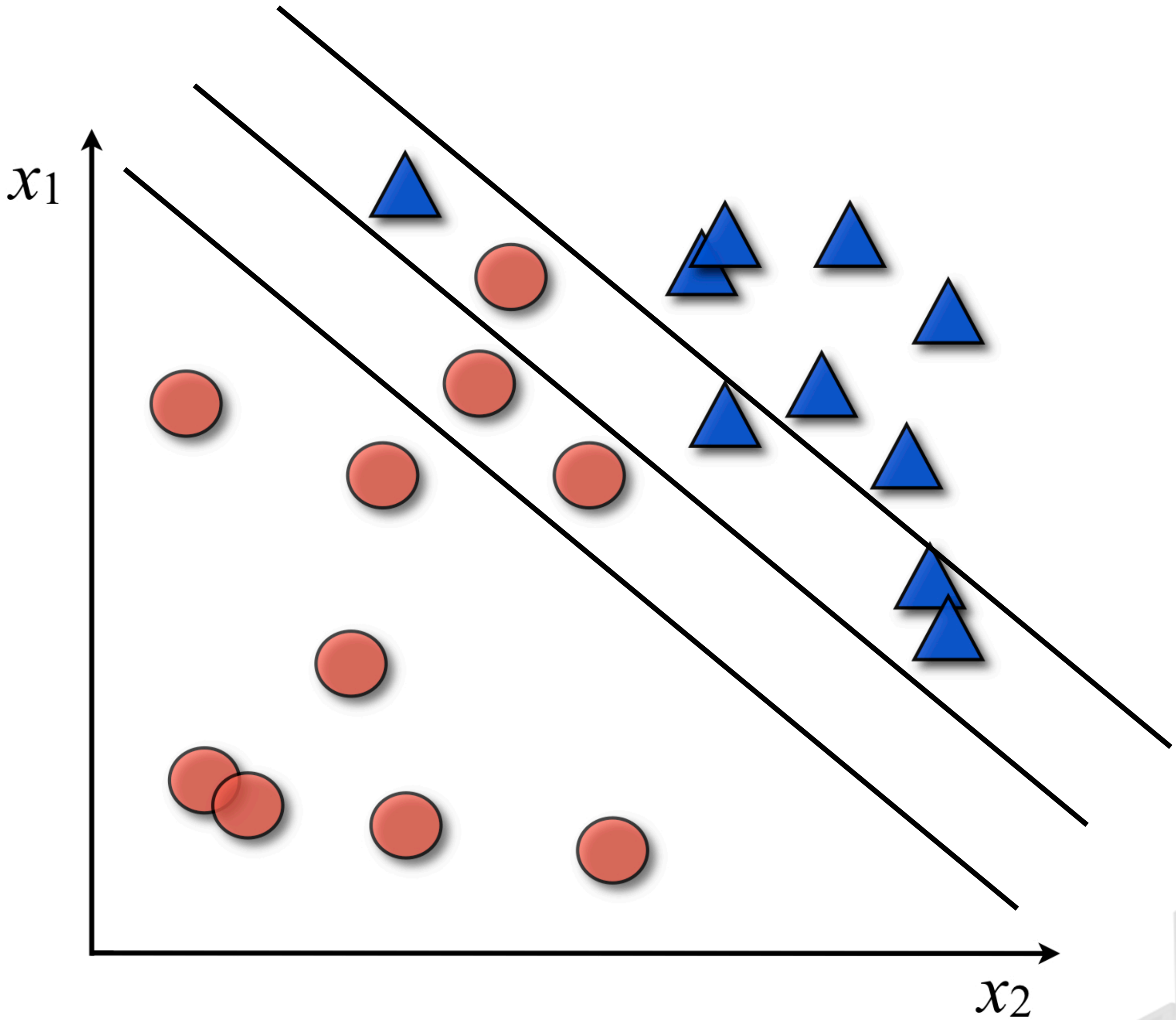
0.478233034571063 he 0.482253521126761
0.253358036127837 an 0.253488372093023
0.298937784522003 re 0.304950495049505
0.155569782330346 ha 0.141408450704225
0.148111332007952 ou 0.126984126984127

Experiments: Poetry

Test	**Function Words** Vector Length	**Function Words** % Misclassified	**Functional Char.-level Bi-grams** Vector Length	**Functional Char.-level Bi-grams** % Misclassified	**Functional Phoneme-level Bi-grams** Vector Length	**Functional Phoneme-level Bi-grams** % Misclassified
Byron vs. Shelley	5	0.185	50	0.1775	20	0.1425
Chapman vs. Shakespeare	5	0.2025	70	0.1650	20	0.1025
Longfellow vs. Coleridge	5	0.0925	20	0.06	20	0.18
Longfellow vs. Poe	5	0.1350	20	0.005	10	0.1550
*Milton vs. Chapman	30	0.0675	70	0.0850	20	0.15
Shelley vs. Keats	20	0.20	-	-	18	0.15
Ovid vs. Vergil	50	0.0950	10	0.0375	-	-

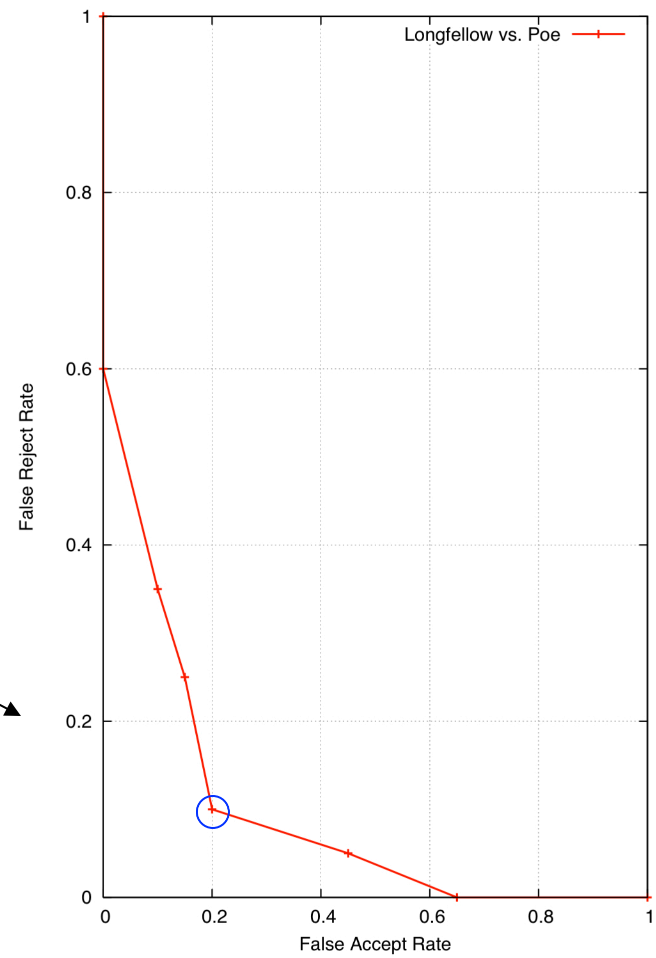
50 training vectors used in all cases except Milton vs. Chapman, which used 100

ROC Analysis



ROC Analysis*

FAR	FRR
Poe Misclassified as Longfellow	**Longfellow Misclassified as Poe**
0.30	0.10



FAR	FRR
Poe Misclassified as Longfellow	**Longfellow Misclassified as Poe**
0.20	0.10

*H. Halteren, “Linguistic Profiling for Author Recognition and Verification,” Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics,” 2004

Post-ROC Analysis: Poetry

Test	**Function Words** Before	**Function Words** After	**Functional Char.-level Bi-grams** Before	**Functional Char.-level Bi-grams** After	**Functional Phoneme-level Bi-grams** Before	**Functional Phoneme-level Bi-grams** After
Byron vs. Shelley	0.185	0.15	0.1775	0.035	0.1425	0.10
Chapman vs. Shakespeare	0.2025	0.165	0.1650	0.0375	0.1025	0.0875
Longfellow vs. Coleridge	0.0925	0.0575	0.06	0.0375	0.18	0.115
Longfellow vs. Poe	0.1350	0.105	0.005	0.0025	0.1550	0.1375
Milton vs. Chapman	0.0675	0.04	0.0850	0.0525	0.15	0.12
Shelley vs. Keats	0.20	0.155	-	-	0.15	0.0725
Ovid vs. Vergil	0.0950	0.0575	0.0375	0.0125	-	-

% Misclassified

The Homeric Question

- What is the provenance of the *Iliad* and *Odyssey*?
- How distinguishable are the poems from one another?
- How heterogeneous is each internally?

The Homeric Question

- "I have assumed the text commented upon is almost entirely Homer's, and the overall cohesiveness has been created by a master storyteller who was usually in full control of his technique."

  — Joseph Russo, Introduction to Od. XVII–XX (Heubeck et. al. 1992, 14)

- "It is now widely accepted that the poem had two main authors: the original poet whom critics call A, and one or more later poets known collectively as B."

  — Manuel Fernández-Galiano, Introduction to Od. XXI (Ibid., 131)

Texts, Samples

	Books ca. 12,000–30,000 chars.	10,000-char samples	5,000-char samples
Iliad	24	57	114
Odyssey	24	41	82
Herodotus' *Histories*	64 samples of 15,000 chars.	96	192

Features

n-grams common to all samples

	n=2	n=3	n=4
5,000	176	115	7
10,000	257	402	66
book	323	926	354

functional n-grams

	n=2	n=3	n=4
5,000	130	110	7
10,000	200	240	40
book	300	430	150

Features

- Character n-grams can obviate the need for parsing in inflected languages*

- Frequent letter combinations pick out the moving parts of words, separating noun- and verb stems from their inflectional endings.

*V. Keselj et al. N-Gram-Based Author Profiles for Authorship Attribution, PACLING 2003

Features

ανδρ captures the noun, “man”

Il. 1.7 ἀνδρῶν	gen. pl.
Il.1.78 ἄνδρα	acc. s.
Il.1.80 ἀνδρὶ	dat. s.
Il.1.146 ἀνδρῶν	gen. pl.
Il.1.151 ἀνδράσιν	dat. pl.
Il.1.172 ἀνδρῶν	gen. pl.
Il.1.242 ἀνδροφόνοιο	compound = "slayer of men"
Il.1.261 ἀνδράσιν	dat. pl.
Il.1.266 ἀνδρῶν	gen. pl.
Il.1.334 ἀνδρῶν	gen. pl.
Il.1.403 ἄνδρες	nom. pl.
Il.1.442 ἀνδρῶν	gen. pl.
Il.1.506 ἀνδρῶν	gen. pl.
Il.1.544 ἀνδρω̂ν	gen. pl.
Il.1.594 ἄνδρες	nom. pl.

Features

οισι captures the dative plural:

Il.1.5 οἰωνοῖσί τε πᾶσι	for all the birds
Il.1.42 σοῖσι βέλεσσιν	by your arrows
Il.1.45 ὤμοισιν	on his shoulders
Il.1.51 αὐτοῖσι	on them themselves
Il.1.58 τοῖσι	among them
Il.1.68 τοῖσι	among them
Il.1.83 ἐν στήθεσσιν ἑοῖσι	in his heart
Il.1.87 Δαναοῖσι	to the Greeks

et. al.

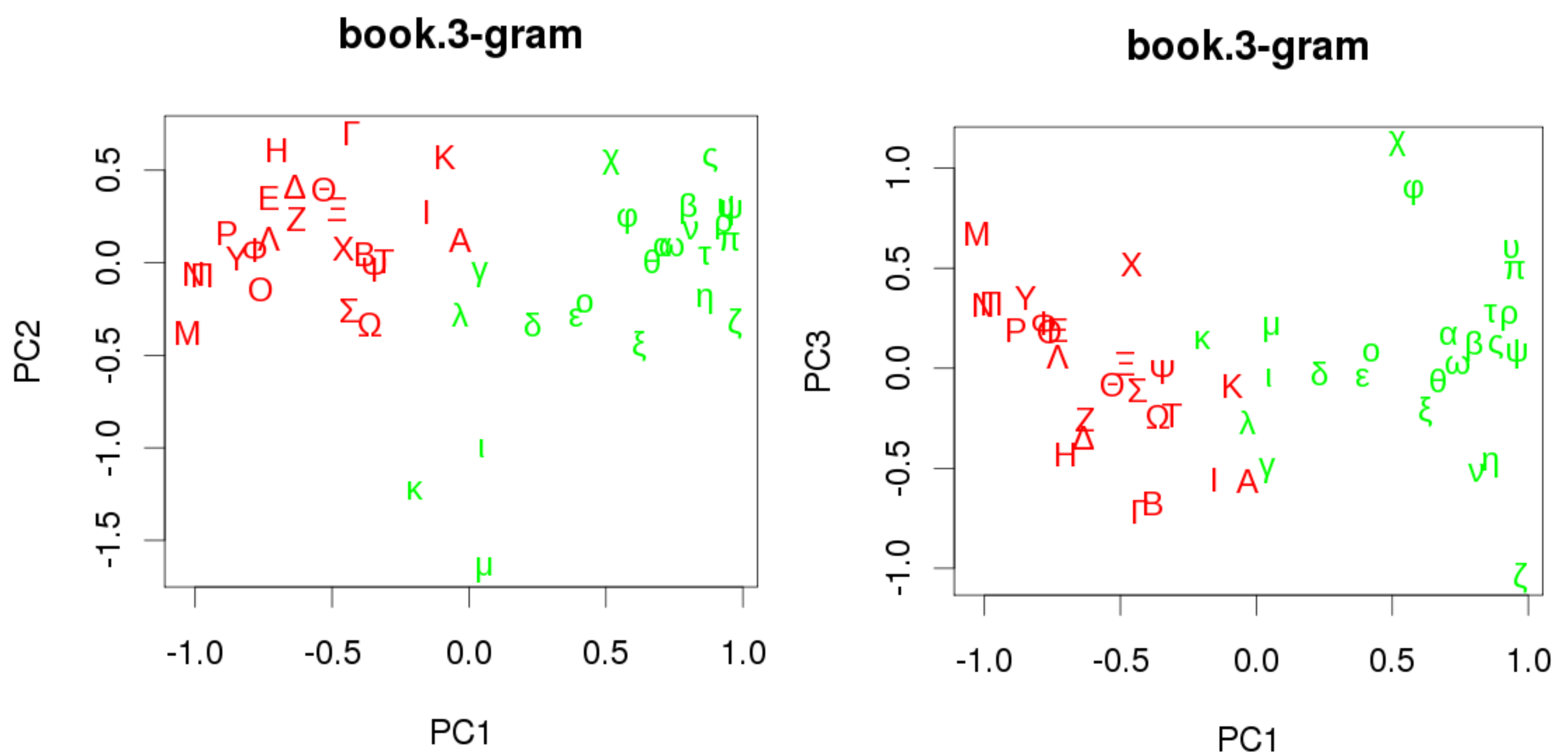
Features

ον	μεν	μενο
αι	και	οισι
εν	οντ	ομεν
οσ	ισι	ισιν
ει	ενο	επει
το	σιν	ενοσ
οι	αλλ	ντεσ
με	αυτ	σθαι
νε	ουσ	οντε
να	ονε	οντο

Classification
success rate

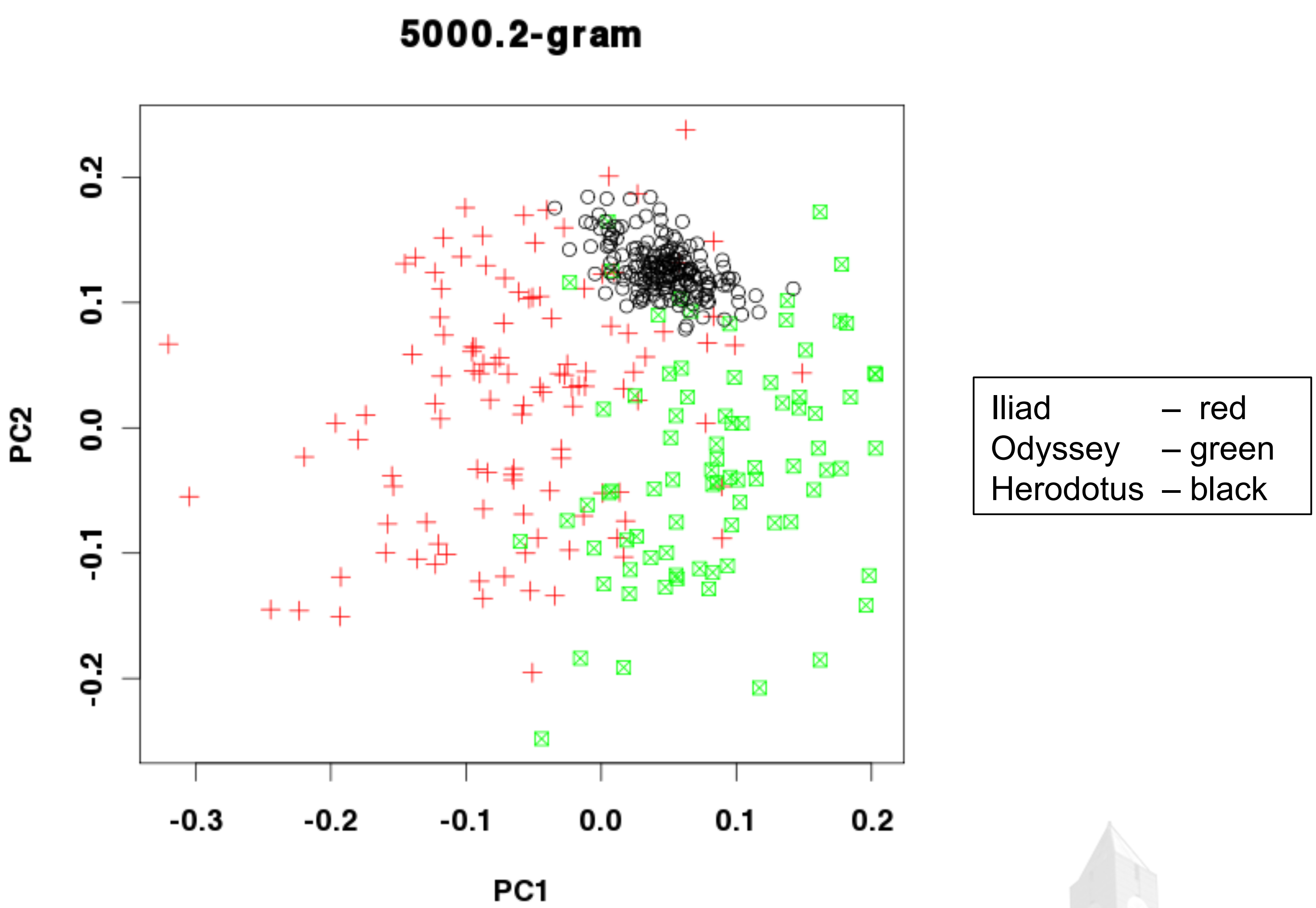
	full feature set			PCA pre-processing			functional feature set		
	n=2	n=3	n=4	n=2	n=3	n=4	n=2	n=3	n=4
5000	88%	87%	58%	87%	82%	57%	89%	87%	58%
10000	81%	95%	70%	94%	98%	73%	81%	98%	73%
book	88%	98%	98%	89%	98%	100%	88%	100%	98%

PCA Plots

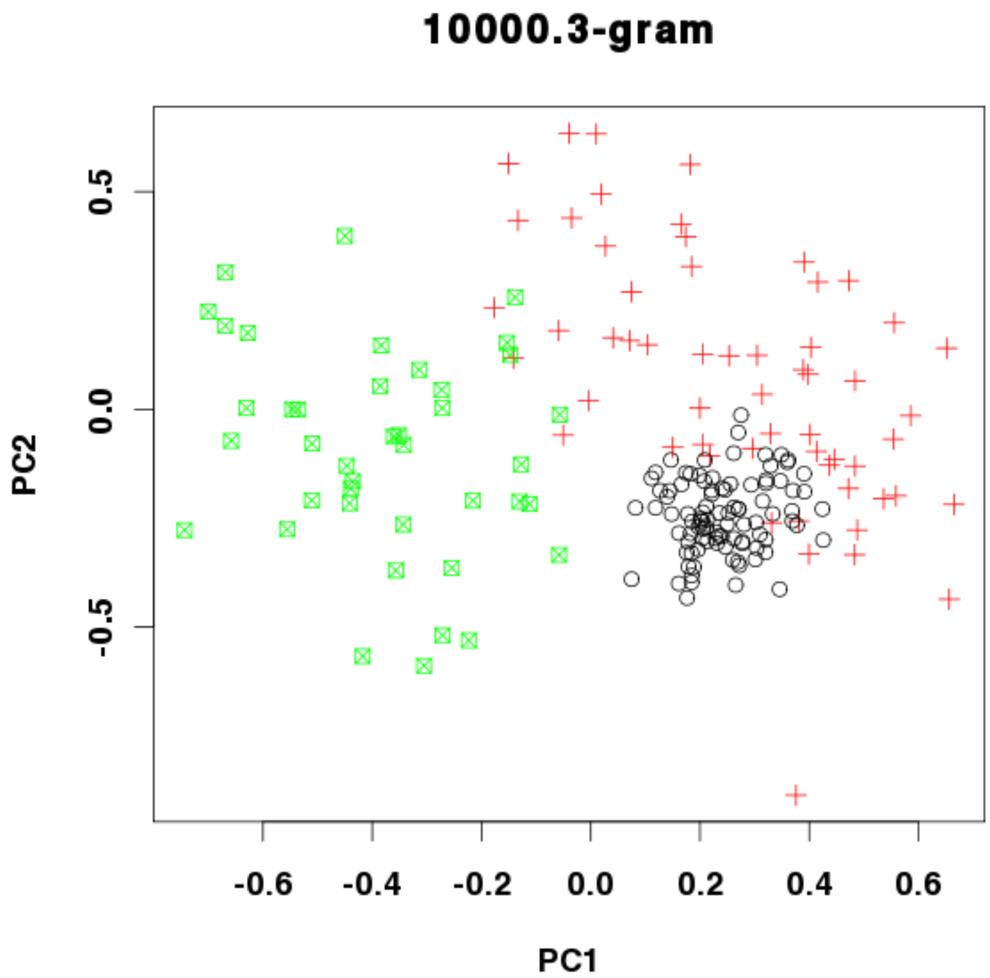


Iliad – red capital letters
Odyssey – green lowercase letters

PCA Plots

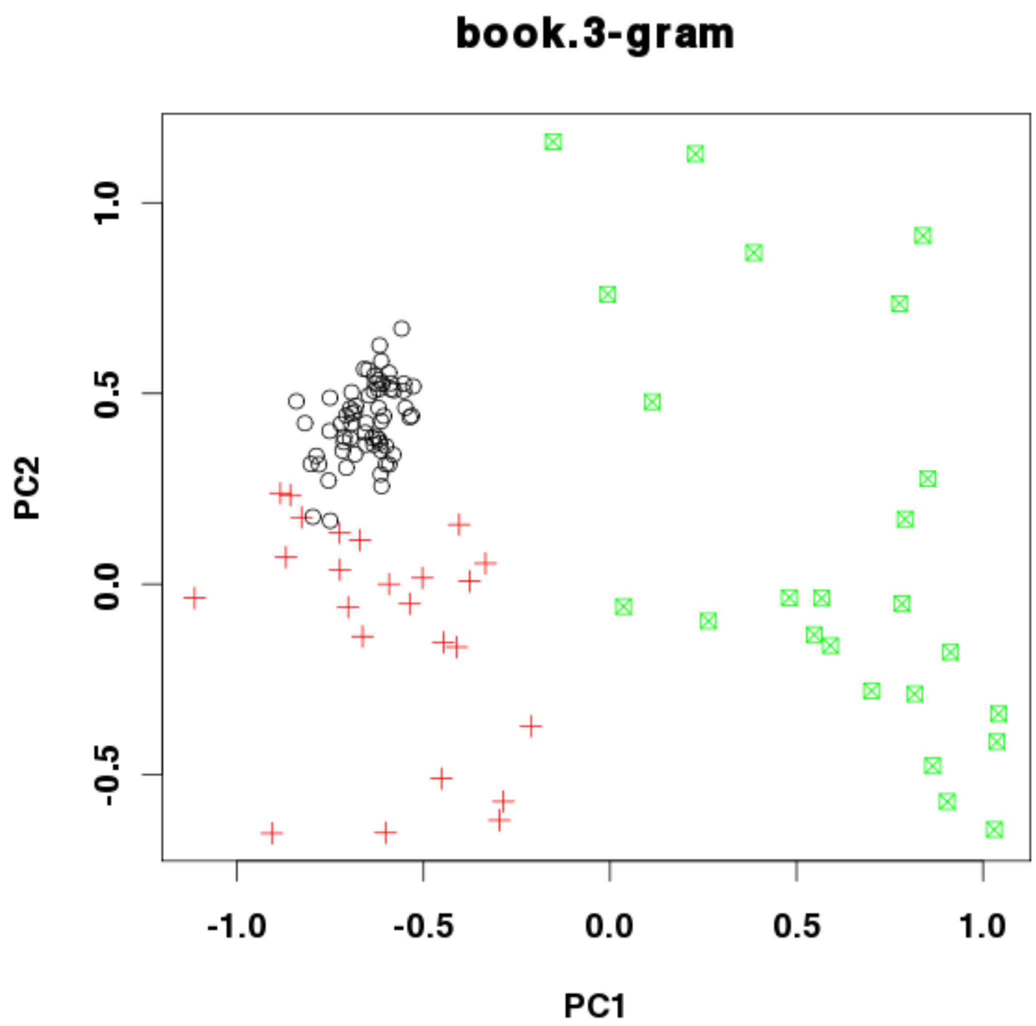

5000.2-gram
PC2
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.2
-0.1
0.0
0.1
0.2
PC1
Iliad – red
Odyssey – green
Herodotus – black

PCA Plots



Iliad	– red
Odyssey	– green
Herodotus	– black

PCA Plots



Iliad – red
Odyssey – green
Herodotus – black

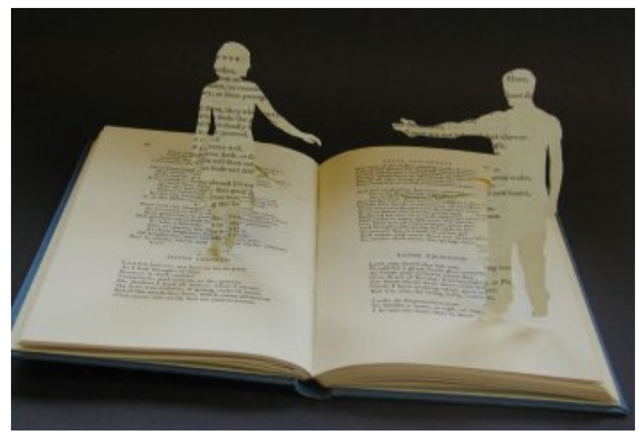
Ongoing Work…

Intertextuality

"Any text is constructed as a mosaic of quotations; any text is the absorption and transformation of another."

The nature of these mosaics is widely varied:

- direct quotations representing a simple and overt intertextuality
- more complex transformations that are intentionally or subconsciously absorbed into a text



Kristeva, "Word, Dialogue, and Novel,"ed. Toril Moi, The Kristeva Reader

New tools in our box

- Functional n-grams apply here, but what about something that is almost opposite of functional?
- Consider elements that occur with lower probabilities:

$$(P_{low} < \Pr(\text{word}_1) < P_{high}) \ \ldots \ (P_{low} < \Pr(\text{word}_2) < P_{high}) \ \ldots \ (P_{low} < \Pr(\text{word}_n) < P_{high})$$

New tools in our box

- How about meter?
  - In practice, the nuance of particular poets, or groups of poets, creates unique variations in meter, giving us a discriminating feature.
    - Add meter information as another dimension to a feature vector for learning
    - Should be useful for group classification

An intriguing text to analyze

- Paul the Deacon's 8th century poem *Angustae Vitae*
  - Strong connection to first-century Neoteric poetry
  - Hypothesis: Paul the Deacon had read Catullus
    - No historical record of this

Some clues…

Catullus II

PASSER, deliciae meae puellae,
quicum ludere, quem in sinu tenere,
cui primum digitum dare appetenti
et acris solet incitare morsus
cum desiderio meo nitenti
carum nescio quid lubet iocari,
credo ut, cum gravis acquiescet ardor,
sit solaciolum sui dolaris,
tecum ludere sicut ipsa possem
et tristis animi levare curas!

Angustae Vitae, lines 1-4:

Angustae vitae fugiunt consortia Musae,
Claustrorum septis nec habitare volunt,
Per rosulenta magis cupiunt sed ludere prata,
Pauperiem fugiunt, deliciasque colunt:

How will it turn out?

- Find out* at DH 2010 in London:
  - http://dh2010.cch.kcl.ac.uk

*Forstall, Jacobson, and Scheirer, “Evidence of Intertextuality: Investigating Paul the Deacon’s *Angustae Vitae,”* to appear at DH 2010

Thank You!

Questions???