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.”

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.”

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 2009¹
  – “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:

\[
P(h \mid t) = \frac{C(th)}{C(t)}
\]

– Generalizing:

\[
P(e_n \mid e_{n-N+1}) = \frac{C(e_{n-N+1}e_n)}{C(e_{n-N+1}^{n-1})}
\]
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\(^2\)

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

<table>
<thead>
<tr>
<th>Test</th>
<th>Function Words Training Vectors</th>
<th>Function Words % Misclassified</th>
<th>Functional Char.-level Bi-grams Training Vectors</th>
<th>Functional Char.-level Bi-grams % Misclassified</th>
<th>Functional Char.-level Tri-grams Training Vectors</th>
<th>Functional Char.-level Tri-grams % Misclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lawrence vs. Austen</td>
<td>90</td>
<td>0.0</td>
<td>100</td>
<td>0.0575</td>
<td>100</td>
<td>0.0275</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>Function Words Training Vectors</th>
<th>Function Words % Misclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lawrence vs. Lawrence</td>
<td>100</td>
<td>0.2125</td>
</tr>
</tbody>
</table>

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
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 twl 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

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

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

# Post-ROC Analysis: Poetry

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

% 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

<table>
<thead>
<tr>
<th></th>
<th>Books ca. 12,000–30,000 chars.</th>
<th>10,000-char samples</th>
<th>5,000-char samples</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Iliad</em></td>
<td>24</td>
<td>57</td>
<td>114</td>
</tr>
<tr>
<td><em>Odyssey</em></td>
<td>24</td>
<td>41</td>
<td>82</td>
</tr>
<tr>
<td>Herodotus' <em>Histories</em></td>
<td>64 samples of 15,000 chars.</td>
<td>96</td>
<td>192</td>
</tr>
</tbody>
</table>
# Features

n-grams common to all samples

<table>
<thead>
<tr>
<th></th>
<th>n=2</th>
<th>n=3</th>
<th>n=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>176</td>
<td>115</td>
<td>7</td>
</tr>
<tr>
<td>10,000</td>
<td>257</td>
<td>402</td>
<td>66</td>
</tr>
<tr>
<td>book</td>
<td>323</td>
<td>926</td>
<td>354</td>
</tr>
</tbody>
</table>

functional n-grams

<table>
<thead>
<tr>
<th></th>
<th>n=2</th>
<th>n=3</th>
<th>n=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>130</td>
<td>110</td>
<td>7</td>
</tr>
<tr>
<td>10,000</td>
<td>200</td>
<td>240</td>
<td>40</td>
</tr>
<tr>
<td>book</td>
<td>300</td>
<td>430</td>
<td>150</td>
</tr>
</tbody>
</table>
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

**αὐδρ** captures the noun, “man”

<table>
<thead>
<tr>
<th>Line</th>
<th>Form</th>
<th>Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>ll. 1.7</td>
<td>ἀνδρῶν</td>
<td>gen. pl.</td>
</tr>
<tr>
<td>ll. 1.78</td>
<td>ἀνδρα</td>
<td>acc. s.</td>
</tr>
<tr>
<td>ll. 1.80</td>
<td>ἀνδρὶ</td>
<td>dat. s.</td>
</tr>
<tr>
<td>ll. 1.146</td>
<td>ἀνδρῶν</td>
<td>gen. pl.</td>
</tr>
<tr>
<td>ll. 1.151</td>
<td>ἀνδράσιν</td>
<td>dat. pl.</td>
</tr>
<tr>
<td>ll. 1.172</td>
<td>ἀνδρῶν</td>
<td>gen. pl.</td>
</tr>
<tr>
<td>ll. 1.242</td>
<td>ἀνδροφόνοιο</td>
<td>compound = &quot;slayer of men&quot;</td>
</tr>
<tr>
<td>ll. 1.261</td>
<td>ἀνδράσιν</td>
<td>dat. pl.</td>
</tr>
<tr>
<td>ll. 1.266</td>
<td>ἀνδρῶν</td>
<td>gen. pl.</td>
</tr>
<tr>
<td>ll. 1.334</td>
<td>ἀνδρῶν</td>
<td>gen. pl.</td>
</tr>
<tr>
<td>ll. 1.403</td>
<td>ἀνδρεῖς</td>
<td>nom. pl.</td>
</tr>
<tr>
<td>ll. 1.442</td>
<td>ἀνδρῶν</td>
<td>gen. pl.</td>
</tr>
<tr>
<td>ll. 1.506</td>
<td>ἀνδρῶν</td>
<td>gen. pl.</td>
</tr>
<tr>
<td>ll. 1.544</td>
<td>ἀνδρῶν</td>
<td>gen. pl.</td>
</tr>
<tr>
<td>ll. 1.594</td>
<td>ἀνδρεῖς</td>
<td>nom. pl.</td>
</tr>
</tbody>
</table>
Features

οῖσι captures the dative plural:

II.1.5 οἶωνοῖσι τε πασί for all the birds
II.1.42 σοῖσι βέλεσσιν by your arrows
II.1.45 μοῖσιν on his shoulders
II.1.51 αὐτοῖσι on them themselves
II.1.58 τοῖσι among them
II.1.68 τοῖσι among them
II.1.83 ἐν στήθεσσιν ἐσοίσι in his heart
II.1.87 Δαναοῖσι to the Greeks
et. al.
## Features

| ον | μὲν | μὲνο | αι | καὶ | οἰσι | 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

<table>
<thead>
<tr>
<th></th>
<th>full feature set</th>
<th>PCA pre-processing</th>
<th>functional feature set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=2</td>
<td>n=3</td>
<td>n=4</td>
</tr>
<tr>
<td>5000</td>
<td>88%</td>
<td>87%</td>
<td>58%</td>
</tr>
<tr>
<td>10000</td>
<td>81%</td>
<td>95%</td>
<td>70%</td>
</tr>
<tr>
<td>book</td>
<td>88%</td>
<td>98%</td>
<td>98%</td>
</tr>
</tbody>
</table>
PCA Plots

book.3-gram

Iliad – red capital letters
Odyssey – green lowercase letters
PCA Plots

5000.2-gram

Iliad – red
Odyssey – green
Herodotus – black
PCA Plots

10000.3-gram

Iliad — red
Odyssey — green
Herodotus — black
PCA Plots

book.3-gram

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

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:

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

Thank You!

Questions???