A Data-Oriented Model of Literary Language

Andreas van Cranenburgh

Rens Bod

HEINRICH HEINE UNIVERSITY DÜSSELDORF

Institut für Sprache und Information
Heinrich Heine University Düsseldorf

Institute for Logic, Language and Computation, University of Amsterdam

EACL, Valencia, April 7, 2017
This talk

Characterizing Literary Language:

- What makes a literary novel *literary*?
- Can a model predict this?
This talk

Characterizing Literary Language:
- What makes a literary novel *literary*?
- Can a model predict this?

Specifically . . .

**Research Question**

are there particular *textual conventions* in literary novels that contribute to readers judging them to be *literary*?
Background

**Definition**

**Literature** is the body of work with the most artistic or imaginative fine writing (Britannica, 1911).
Background

**Definition**

*Literature* is the body of work with the most artistic or imaginative fine writing (Britannica, 1911).

- Demarcation problem
- Some argue text is irrelevant, only context/prestige matters

- Therefore, interesting to *quantify* influence of text
- **NB:** not the same as success, popularity, quality, &c.
The Riddle of Literary Quality

Corpus:
- 401 recent Dutch novels (translated & original)
- Published 2007–2012
- Selected by popularity

http://www.literaryquality.huygens.knaw.nl
The Riddle of Literary Quality

Corpus:
- 401 recent Dutch novels (translated & original)
- Published 2007–2012
- Selected by popularity

Contrast: Gutenberg, Google Books
- more books (thousands, millions)
- not representative (volunteer work, digital availability)
- not contemporary (19th century)

cf. Pechenick et al. (2015), PloS ONE. Characterizing the Google Books Corpus: Strong Limits (…)
Survey ratings: 401 novels; N=14k

- Definitely not literary
- Not literary
- Tending towards non-literary
- Bordering literary and non-literary
- Somewhat literary
- Literary
- Highly literary

http://www.hetnationalelezersonderzoek.nl
Survey ratings: 401 novels; N=14k

![Chart showing survey ratings of novels, with categories ranging from Definitely not literary to Highly literary. The chart also displays a graph with mean ratings and 95% confidence intervals.](http://www.hetnationalelezersonderzoek.nl)
Survey ratings: 401 novels; N=14k

Definitely not literary  Tending towards non-literary  Bordering literary and non-literary  Somewhat literary  Literary  Highly literary

Constraints:
▶ ≥ 50 ratings
▶ ≥ 2000 sent.
369 novels remain
91 % novels conf. int. < 0.5

http://www.hetnationalelezersonderzoek.nl
### Overview

The document-feature matrix is used for predicting survey ratings. It contains features such as "sent.len", "BoW", and "genre". For example, the matrix includes entries for "50 shades of grey", "eat pray love", and "super high brow stuff" with corresponding ratings like 9.1, 17.9, and 14.1. The task is to predict the survey rating of these novels based on their features.
Experimental setup

Task: predict mean literary rating (1–7)
Training data: 1000 sentences per novel
Evaluation metric: $R^2$ ($\approx$ % variation explained, baseline=0.0, perfect=100 %)
Show incremental improvement with each type of feature.
Simple Stylistic Measures

\[ R^2 \]

Mean sent. len.
+ % Direct speech
+ % Basic vocab. (top 3000 words)
+ Compression ratio (bzip2)
+ Cliche expressions

Table: Basic features
Simple Stylistic Measures

<table>
<thead>
<tr>
<th>Feature</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean sent. len.</td>
<td>16.4</td>
</tr>
<tr>
<td>+ % Direct speech</td>
<td>23.1</td>
</tr>
<tr>
<td>+ % Basic vocab. (top 3000 words)</td>
<td>23.5</td>
</tr>
<tr>
<td>+ Compression ratio (bzip2)</td>
<td>24.4</td>
</tr>
<tr>
<td>+ Cliche expressions</td>
<td>30.0</td>
</tr>
</tbody>
</table>

**Table**: Basic features, incremental scores.
Strong lexical baselines

Setup: Linear Support Vector Regression, 5-fold crossvalidation

$R^2$

Basic features
+ LDA: 50 topic weights
+ Word bigrams
+ Char. 4-grams

On average, $59.9\%$ of variation in ratings ($R^2$) is explained using basic and lexical features. The prediction is off by $0.64$ (RMSE) out of $0–7$. 
Strong lexical baselines

Setup: Linear Support Vector Regression, 5-fold crossvalidation

<table>
<thead>
<tr>
<th>Features</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic features</td>
<td>30.0</td>
</tr>
<tr>
<td>+ LDA: 50 topic weights</td>
<td>52.2</td>
</tr>
<tr>
<td>+ Word bigrams</td>
<td>59.5</td>
</tr>
<tr>
<td>+ Char. 4-grams</td>
<td>59.9</td>
</tr>
</tbody>
</table>

On average,

- 59.9% of variation in ratings ($R^2$) is explained using basic and lexical features.
- the prediction is off by 0.64 (RMSE) out of 0–7.
n-gram limitations

1. fixed $n$:
   no MWE, long-distance relations
2. no linguistic abstraction:
   e.g., syntactic categories, grammatical functions
3. small features:
   harder to interpret
1. fixed $n$:  
   no MWE, long-distance relations

2. no linguistic abstraction:  
   e.g., syntactic categories, grammatical functions

3. small features:  
   harder to interpret

   ▶ Larger features $\Rightarrow$ combinatorial explosion
   ▶ Use data-driven feature selection
Recurring Tree Fragments

- Syntactic tree fragments of arbitrary size (connected subsets of tree productions)
- Extract automatically from training data: find overlapping parts of parse trees
- Apply cross-validation
- Feature selection using correlation with literary rating

```
fold 1
fold 2
fold 3
fold 4
fold 5
```
Example fragments

Mean literary rating vs. fragment count for different structures:
- **NP-obj1**: $r = 0.526$
- **ROOT**: $r = -0.417$
- **PP-mod**: $r = 0.4$
### Results w/Fragments

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic features</td>
<td>30.0</td>
</tr>
<tr>
<td>+ LDA: 50 topic weights</td>
<td>52.2</td>
</tr>
<tr>
<td>+ Word bigrams</td>
<td>59.5</td>
</tr>
<tr>
<td>+ Char. 4-grams</td>
<td>59.9</td>
</tr>
<tr>
<td>+ Syntactic fragments</td>
<td>62.2</td>
</tr>
</tbody>
</table>
Results w/Fragments

\[
\begin{array}{ll}
\text{Basic features} & 30.0 \\
+ \text{LDA: 50 topic weights} & 52.2 \\
+ \text{Word bigrams} & 59.5 \\
+ \text{Char. 4-grams} & 59.9 \\
+ \text{Syntactic fragments} & 62.2 \\
\end{array}
\]

- Syntax gives modest performance improvement
- However, features are linguistically more interesting
Analysis of tree fragments

Fragments positively correlated w/literary ratings:
► Many small fragments
► Indicators of more complex syntax, e.g.:

appositive NPs:
      His name was Adrian Finn, a tall, shy boy who (…)
      (Barnes, Sense of an ending)

complex, nested NPs/PPs:
      (…) a whole storetank of existential rage
      (Barnes, Sense of an ending)

discontinuous constituents:
      ‘Miss Aibagawa,’ declared Ogawa, ‘is a midwife.’
      (Mitchell, Thousand autumns of J. Zoet)
Metadata

**Coarse genre:** Fiction, Suspense, Romance, Other

**Translated vs. originally Dutch**

**Author gender:** male, female, mixed/unknown
Metadata

Coarse genre: Fiction, Suspense, Romance, Other
Translated vs. originally Dutch
Author gender: male, female, mixed/unknown

<table>
<thead>
<tr>
<th>Basic features</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Auto. induced feat.</td>
<td>61.2</td>
</tr>
<tr>
<td>+ Genre</td>
<td>74.3</td>
</tr>
<tr>
<td>+ Translated</td>
<td>74.0</td>
</tr>
<tr>
<td>+ Author gender</td>
<td>76.0</td>
</tr>
</tbody>
</table>

Table: Metadata features; incremental scores.
Prediction scatter plot

- Fiction
- Suspense
- Romantic
- Other

Actual reader judgments:
- James
- Fifty shades of Grey
- Kinsella
- Remember me?
- Smeets
- Afrekening
- Gilbert
- Eat pray love
- French
- Blue monday
- Stockett
- The help
- Donoghue
- Room
- Franzen
- Freedom
- Barnes
- Sense of an ending
- Lewinsky
- Johannistag
- Mortier
- Godenslaap
- Rosenboom
- Zoete mond
- Baldacci
- Hell’s corner
- French
- Blue monday
- Fiction
- Suspense
- Romantic
- Other
Research Question

are there particular textual conventions in literary novels that contribute to readers judging them to be literary?

- Yes! Literary conventions are non-arbitrary because they are associated with textual features
- Literariness can be predicted from text to a large extent: text-intrinsic literariness
Conclusion

Research Question

are there particular textual conventions in literary novels that contribute to readers judging them to be literary?

- Yes! Literary conventions are non-arbitrary because they are associated with textual features
- Literariness can be predicted from text to a large extent: text-intrinsic literariness
- Cumulative improvements with ensemble of features
- Robust result: both coarse & fine rating differences are predicted
- Literature is characterized by a larger inventory of lexico-syntactic constructions
THE END

Dissertation & code: http://andreasvc.github.io

Figure: Huff (1954). How to lie with statistics.
BUT WAIT, THERE’S MORE
Fragment size (non-terminals)

![Diagram showing the distribution of fragment sizes with positive and negative correlation.](image-url)
Syntactic category of root node

The graph shows the number of fragments for different syntactic categories of the root node, with positive and negative correlations indicated by bars colored white and gray, respectively.
Function tag of root node

![Graph showing number of fragments for different function tags. The x-axis represents function tags ranging from `mod` to `det`, and the y-axis shows the number of fragments. The graph compares positive and negative correlations.](attachment:image.png)
1. n-hd, r=0.52
2. NP-su SMAIN-dp, SMAIN-dp r=0.46
3. lid-det n-hd r=0.42
4. lid-det NP-app r=0.41
5. SMAIN-dp DU r=0.41
6. vz-hd CONJ-obj1 NP-obj1 r=0.41
7. ww-hd NP-su r=0.41
8. lid-det n-hd r=0.41
9. (SMAIN-dp ... ...) r=0.41
10. ln r=0.41
7770. ?        r=-0.32
7771. ' tsw-tag DU .    r=-0.33
7772. NP-su      r=-0.34
7773. vnv-hd     r=-0.34
7774. echt       r=-0.34
7775. Oké        r=-0.34
7776. ' Ik SMAIN .   r=-0.35
7777. ' DU .        r=-0.39
7778. ' NP-su SMAIN .   r=-0.40
7779. ww-hd adj-mod   r=-0.43