A Data-Oriented Model of Literary Language

Andreas van Cranenburgh
Rens Bod

HEINRICH HEINE
UNIVERSITÄT 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

The Riddle of Literary Quality

Corpus:

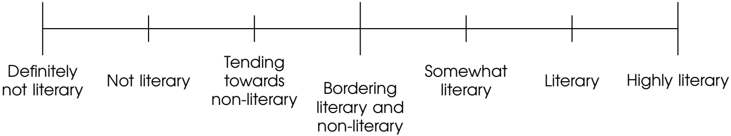
- 401 recent Dutch novels (translated & original)
- Published 2007–2012
- Selected by popularity

Contrast: Gutenberg, Google Books

- more books (thousands, millions)
- not representative (volunteer work, digital availability)
- not contemporary (19th century)

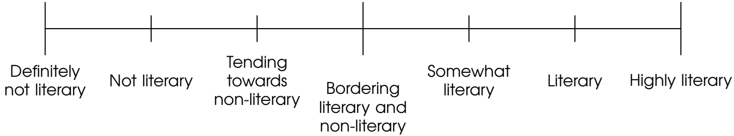
cf. Pechenick et al. (2015), PloS ONE. Characterizing the Google Books Corpus: Strong Limits (...)

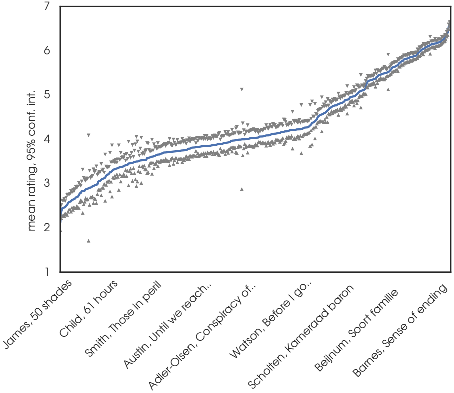
Survey ratings: 401 novels; N=14k



http://www.hetnationalelezersonderzoek.nl

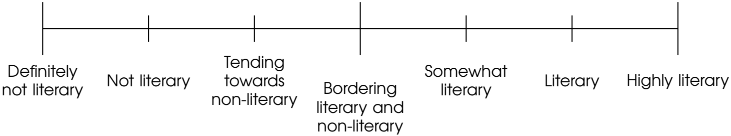
Survey ratings: 401 novels; N=14k

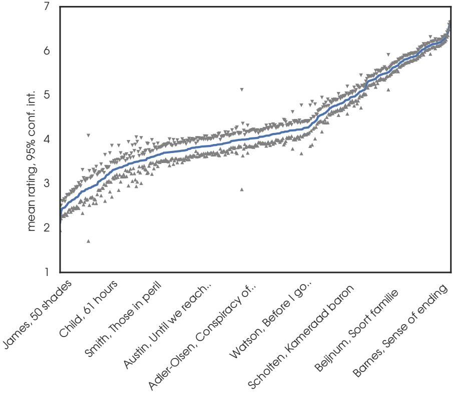




http://www.hetnationalelezersonderzoek.nl

Survey ratings: 401 novels; N=14k





Constraints:

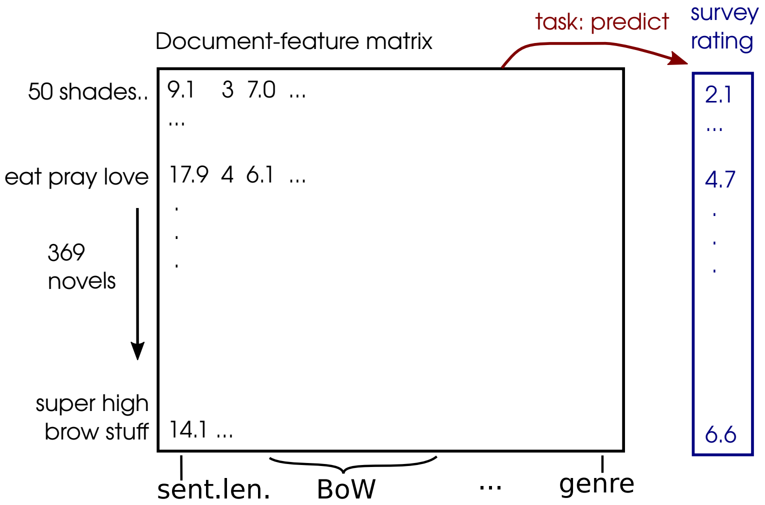
- $\geq$ 50 ratings
- $\geq$ 2000 sent.

369 novels remain

91 % novels conf. int. < 0.5

http://www.hetnationalelezersonderzoek.nl

Overview


Document-feature matrix
task: predict
survey
rating
50 shades..
eat pray love
369
novels
super high
brow stuff
9.1 3 7.0 ...
...
17.9 4 6.1 ...
14.1 ...
2.1
...
4.7
6.6
sent.len.
BoW
...
genre

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

	R^2
Mean sent. len.	16.4
+ % Direct speech	23.1
+ % Basic vocab. (top 3000 words)	23.5
+ Compression ratio (bzip2)	24.4
+ Cliche expressions	30.0

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	

Strong lexical baselines

Setup: Linear Support Vector Regression,
5-fold crossvalidation

	R^2
Basic features	30.0
+ LDA: 50 topic weights	52.2
+ Word bigrams	59.5
+ Char. 4-grams	59.9

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

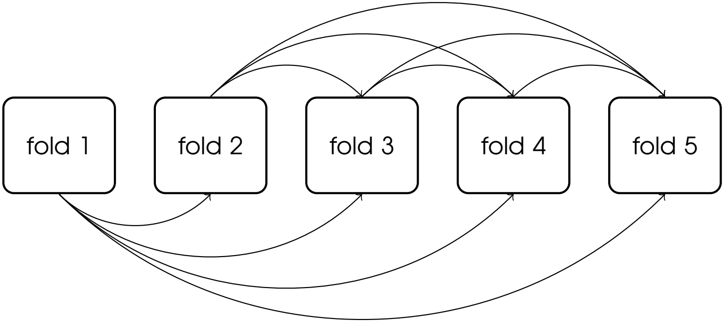
n-gram limitations

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

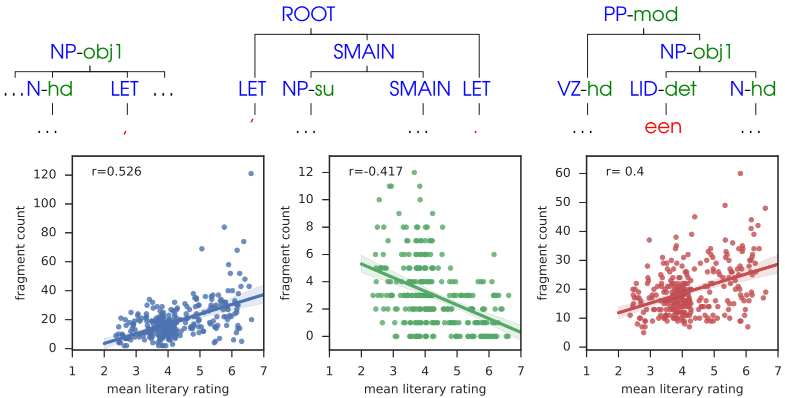
- Larger features $\Rightarrow$ combinatorial explosion
- Use data-driven feature selection

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



Example fragments



Results w/Fragments

	R^2
Basic features	30.0
+ LDA: 50 topic weights	52.2
+ Word bigrams	59.5
+ Char. 4-grams	59.9
+ Syntactic fragments	62.2

Results w/Fragments

	R^2
Basic features	30.0
+ LDA: 50 topic weights	52.2
+ Word bigrams	59.5
+ Char. 4-grams	59.9
+ Syntactic fragments	62.2

- 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

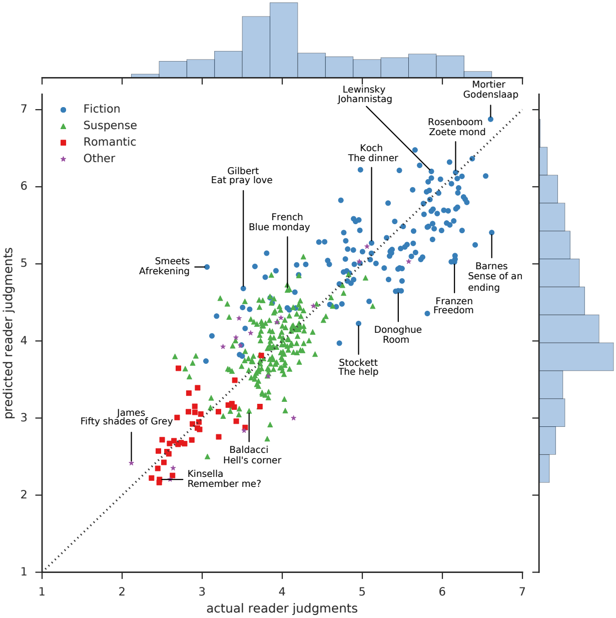
Translated vs. originally Dutch

Author gender: male, female, mixed/unknown

	R^2
Basic features	30.0
+ Auto. induced feat.	61.2
+ Genre	74.3
+ Translated	74.0
+ Author gender	76.0

Table: Metadata features; incremental scores.

Prediction scatter plot



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

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

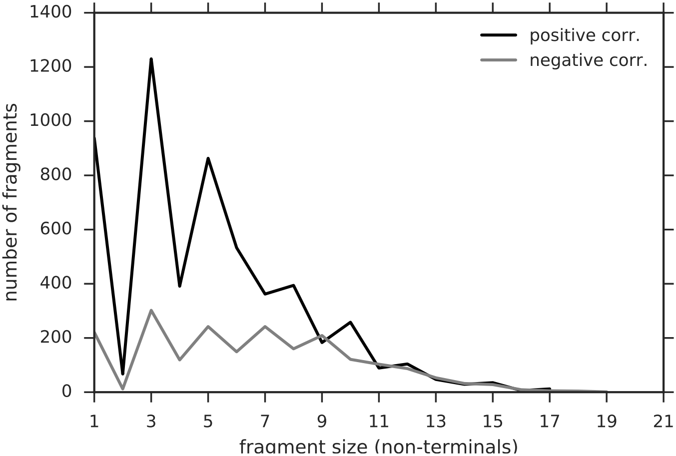
THE END

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

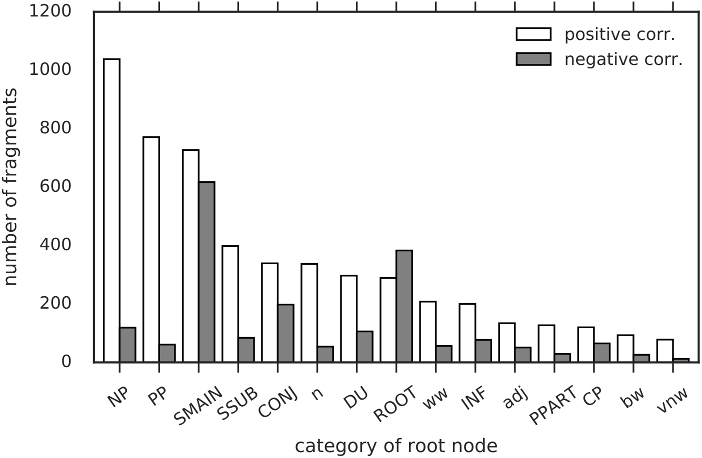
Figure: Huff (1954). How to lie with statistics.

BUT WAIT, THERE'S MORE

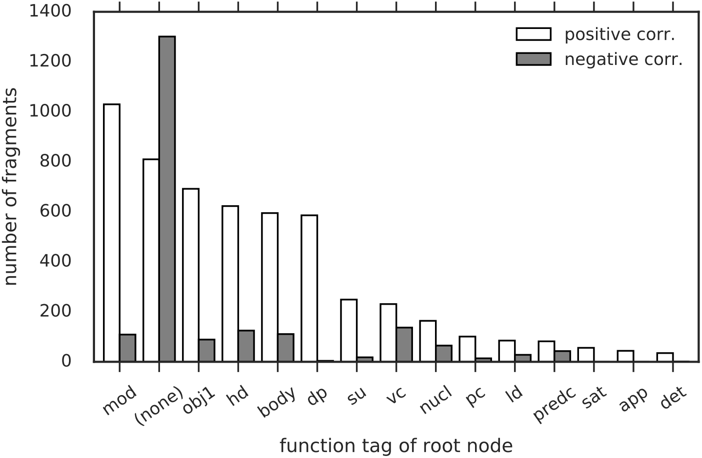
Fragment size (non-terminals)



Syntactic category of root node



Function tag of root node



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. In   r=0.41

7770. ? r=-0.32
7771. ’ tsw-tag DU . r=-0.33
7772. NP-su r=-0.34
7773. vnw-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