Markers of Literary Language—
A Computational-Linguistic Odyssey

Andreas van Cranenburgh

Huygens ING
Royal Netherlands Academy of Arts and Sciences

Institute for Logic, Language and Computation
University of Amsterdam

January 15, 2016
Background

**Definition**

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

Other definitions . . .

**Formalist definition:** literary language distinguishes itself from standard language through foregrounding, defamiliarization.

**Social definition:** critics and publishers determine status of novels

**Computational Humanities?** investigate true nature of literary conventions using computational tools.
Heumakers (2015): *De esthetische revolutie*

Modern, Western notion of literature:
- Late 18th century invention, within *Romanticism*
- Important features:
  - autonomous
  - originality
  - not just entertainment
  - not limited by genre
  - social criticism
The Riddle of Literary Quality

Corpus:

- 401 recent Dutch novels
- Published 2007–2012
- Selected by popularity

Survey:

- For books read:
  Likert scales (1–7) how literary, good?
- about 14,000 readers completed the survey
Simple Stylistic Measures

- Words per Sentence (WPS)
- Modifier Constituents (MOD)
- Direct Speech (DS)
- Common Vocabulary (CV)
Simple Stylistic Measures

- Words per Sentence (WPS)
- Modifier Constituents (MOD)
- Direct Speech (DS)
- Common Vocabulary (CV)

<table>
<thead>
<tr>
<th>Literariness</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPS</td>
<td>0.39*</td>
</tr>
<tr>
<td>MOD</td>
<td>0.27*</td>
</tr>
<tr>
<td>DS</td>
<td>-0.39*</td>
</tr>
<tr>
<td>CV</td>
<td>-0.31*</td>
</tr>
</tbody>
</table>

* means significant correlation with $p < 0.001$
Clichés

Definition

A cliché expression is a fixed, conventionalized yet compositional multi-word expression which has become overused to the point of losing its original meaning or effect.

- (Kijk,) dat bedoel ik nou. (Look,) that’s what I mean.
- Geen (bier) meer voor jou! No more (beer) for you!
- Y, zoals X dan zou zeggen. Y, as X would say.
- Daar zit een boek/artikel in! That’s material for a book/paper!
Results with Cliches

Pearson correlation: $r = -0.48$, $p = 8.9 \times 10^{-24}$
Fragments

- Data-Oriented Parsing (Scha 1990, Bod 1992)
- Language use is memory-based, depending on arbitrary fragments of language experience
- Syntactic tree fragments of arbitrary size (connected subsets of tree productions)
- Extract automatically from corpus
Fragments

ROOT

SMAIN

LET NP-su SMAIN LET

NP-su

LID-det N-hd

WW(pv)-hd ADJ(vrij)-mod

mean literariness rating

fragment count

mean literariness rating

fragment count

mean literariness rating

fragment count
1. Extract common fragments
2. Apply frequency thresholds:
   \( \geq 50 \) occurrences in \( \geq 5\% \) of texts
3. Apply correlation threshold:
   correlated s.t. \( p < 0.05 \)
4. Remove redundant fragments:
   discard fragments correlated with \( |r| > 0.5 \)
actual reader judgments
1 2 3 4 5 6 7
predicted reader judgments
r2_score = 0.54
Fragments

<table>
<thead>
<tr>
<th></th>
<th>RMS error</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragments</td>
<td>0.68</td>
<td>54.2</td>
</tr>
<tr>
<td>Bigrams</td>
<td>0.66</td>
<td>56.7</td>
</tr>
<tr>
<td>Combined</td>
<td>0.66</td>
<td>56.9</td>
</tr>
<tr>
<td>Interpolated</td>
<td>0.65</td>
<td>57.3</td>
</tr>
</tbody>
</table>
Conclusion

- Yes, literary conventions are related to textual features
- Literariness can be predicted from text to quite a large extent
- We presented a data-oriented approach with fully parsed sentences that improves on simpler baselines
THE END

Don't be a novelist—be a statistician, much more scope for the imagination.