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

ILLC Midwinter Colloquium 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.

Background

Heumakers (2015): *De esthetische revolutie*

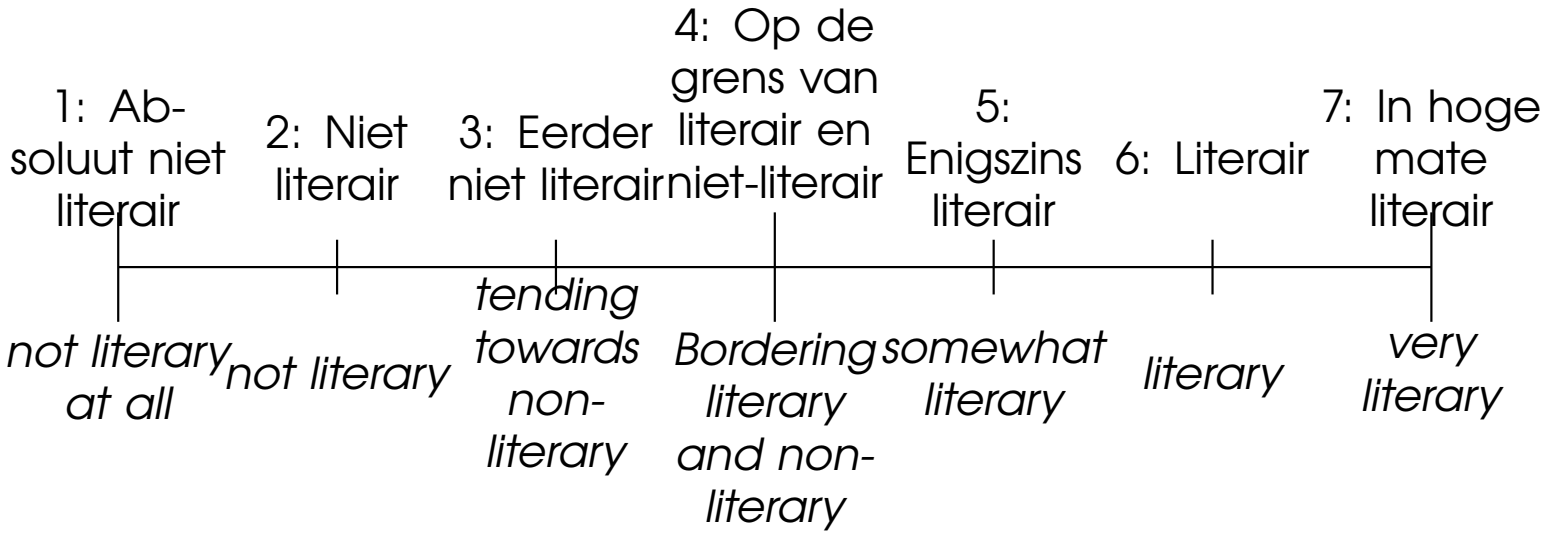
Modern, Western notion of literature:

- 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)

	Literariness	Quality
WPS	0.39*	0.24*
MOD	0.27*	0.02
DS	-0.39*	-0.04
CV	-0.31*	-0.15

* means significant correlation with $p < 0.001$

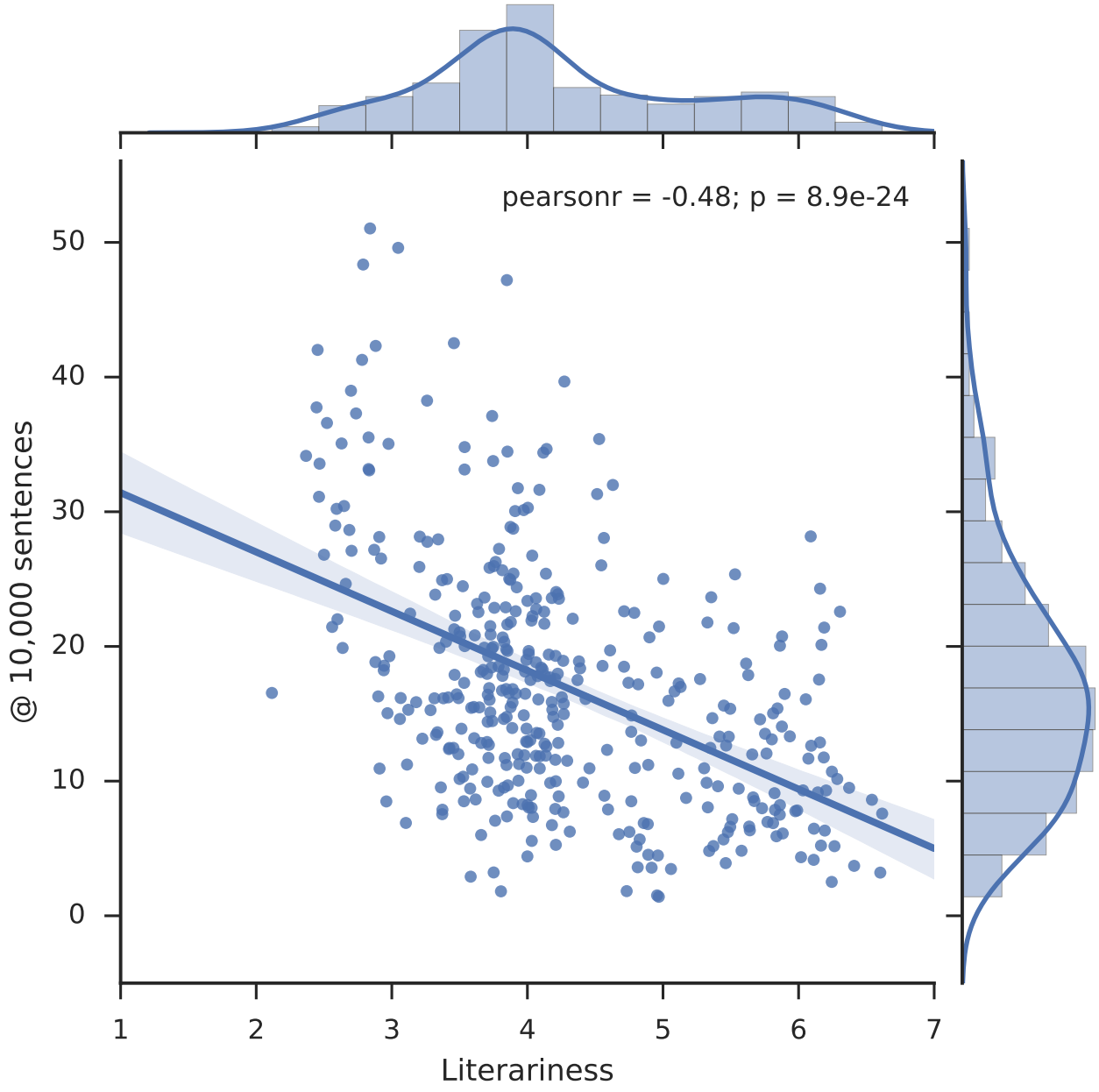
Clichés

Definition

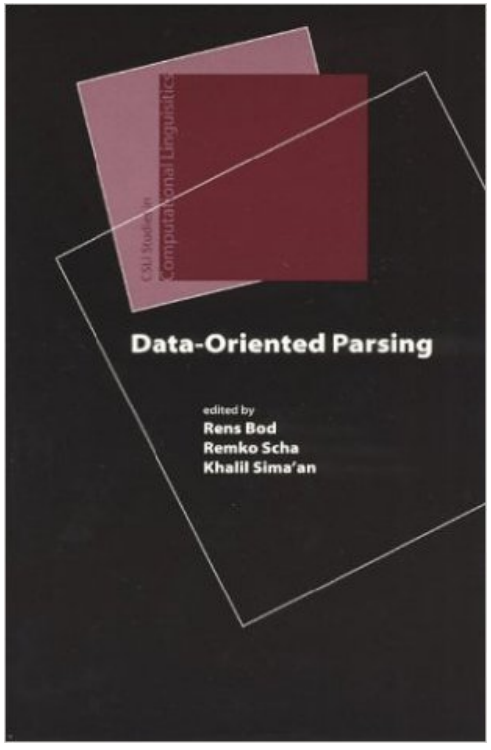
A cliché expression is a fixed, conventionalized yet compositional multi-word expression which has become overused to the point of losing its original meaning or effect.

- (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

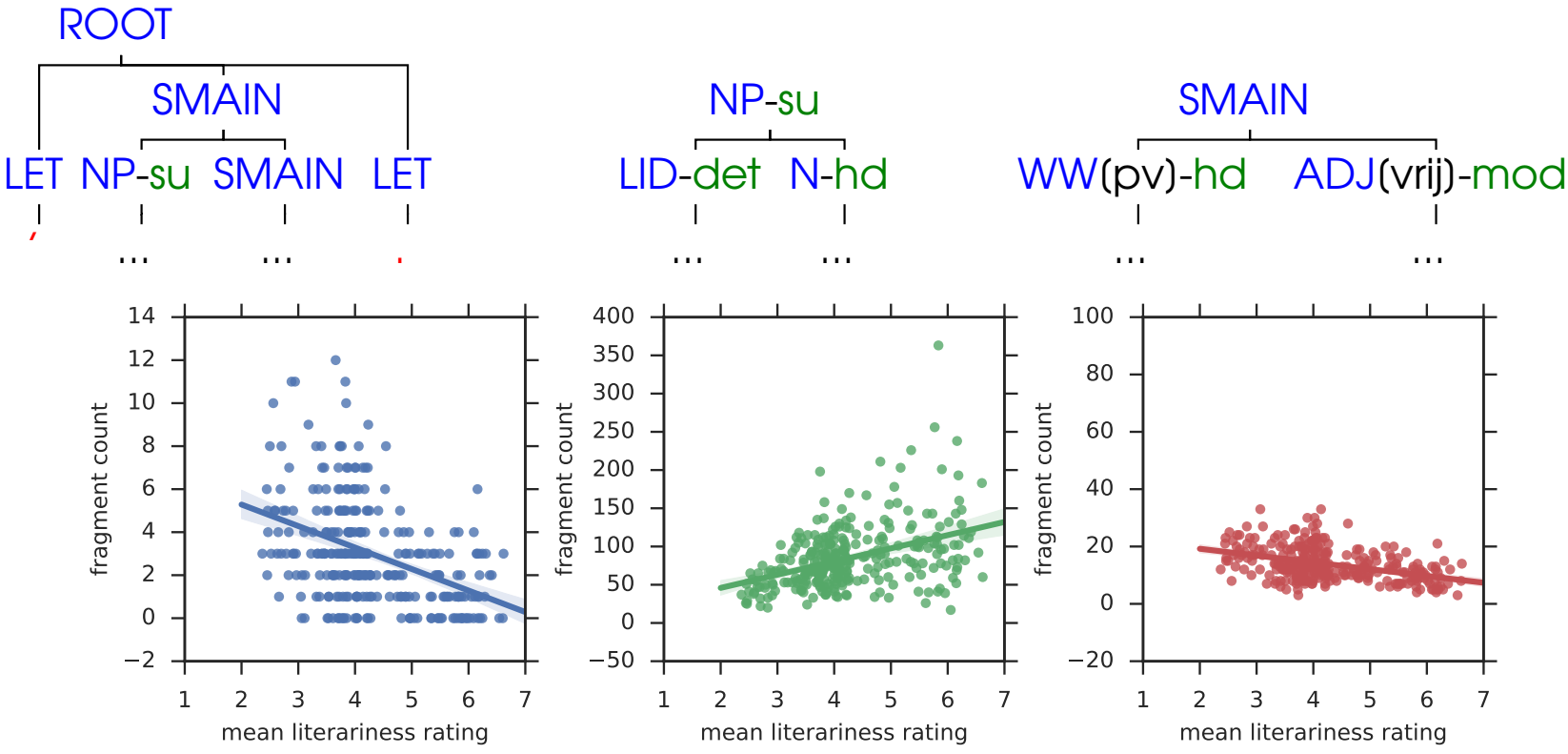


Fragments

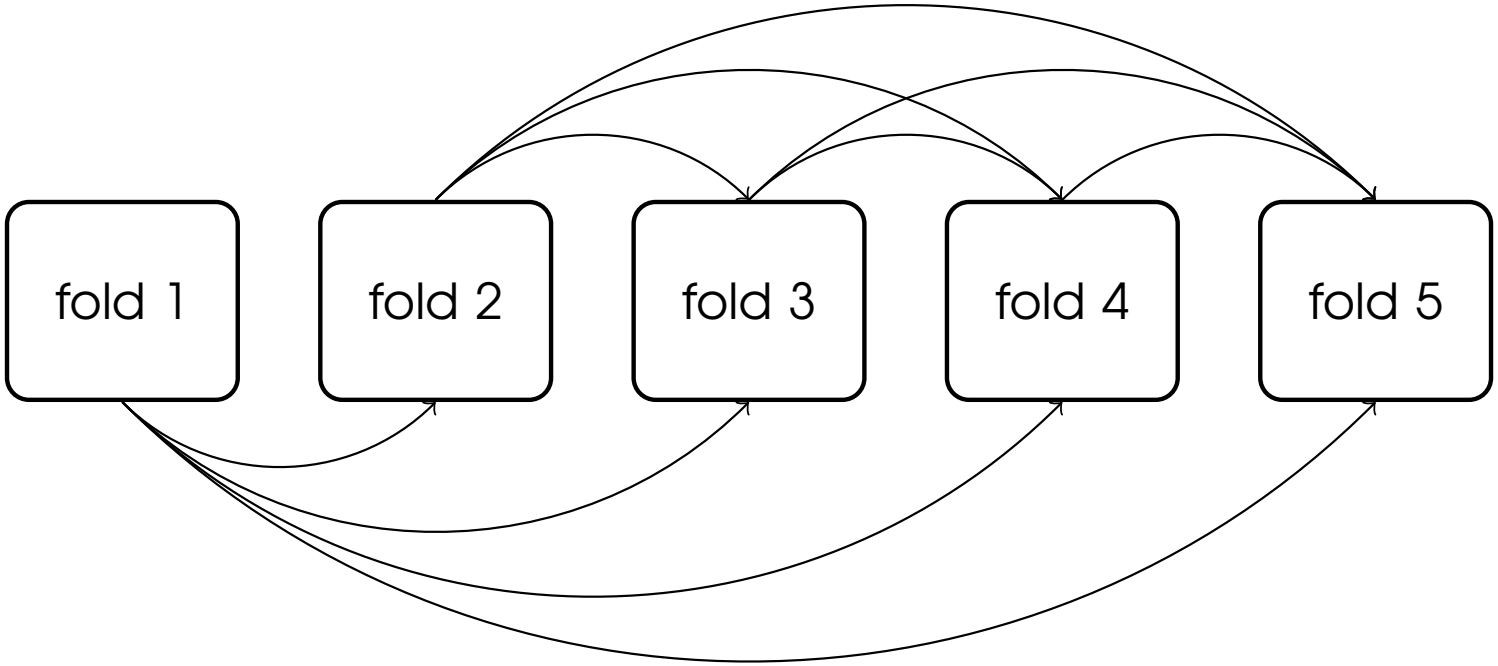


- Data-Oriented Parsing
  (Scha 1990, Bod 1992)
- Language use is memory-based, depending on arbitrary fragments of language experience
- Syntactic tree fragments of arbitrary size
  (connected subsets of tree productions)
- Extract automatically from corpus

Fragments

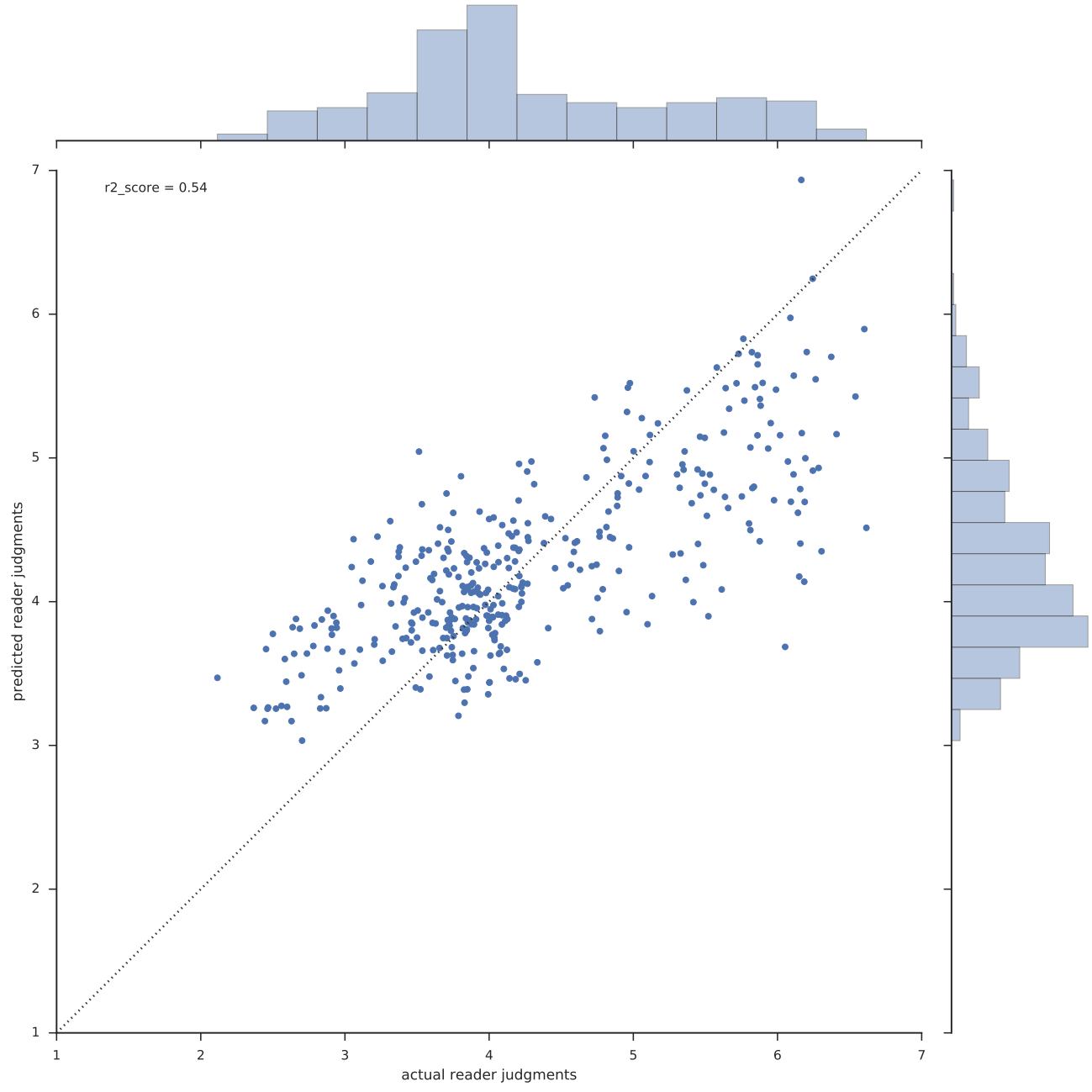

ROOT
SMAIN
LET NP-su SMAIN LET
NP-su
LID-det N-hd
SMAIN
WW(pv)-hd ADJ(vrij)-mod
fragment count
mean literariness rating
fragment count
mean literariness rating
fragment count
mean literariness rating

Fragments



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 $|r| > 0.5$

Fragments


r2_score = 0.54
predicted reader judgments
actual reader judgments
1
2
3
4
5
6
7

Fragments

	RMS error	R^2
Fragments	0.68	54.2
Bigrams	0.66	56.7
Combined	0.66	56.9
Interpolated	**0.65**	**57.3**

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...