# Cultural Analytics

## ENGL 64.05

Fall 2019

Prof. James E. Dobson

October 9, 2019

# Aristotelian Plot

- Narrative stories are “wholes” and require a plot to make them complete.
- The whole is composed of a beginning, middle, and end.
- The ending makes the plot complete by wrapping up within the narrative logic.

| Sjuzet / Plot | Fabula / Story |
| --- | --- |
| • The formal arrangement of the narrated events as they are presented to the reader.<br>• Not necessarily the order in which they events take place within the logic of the story. | • The material of the narrative, the events themselves as they happen in chronological order.<br>• Not necessarily in the same order as the unfolding of these events in narrative time. |

# Book Segmentation

- Segmentation enables
  - comparison across books of uneven length
  - comparison within a book
  - a sense of temporal change
  - a sense of development
  - a way to access the sjuzet or plot of a narrative

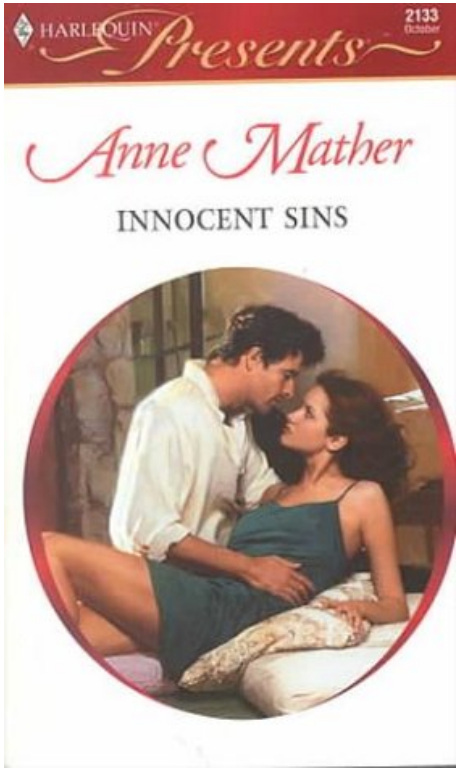
- What are the best units of comparison for books?
  - Entire book?
  - Chapters?
  - Paragraphs?
  - Uniform segments?

# Proxies for Plot?

- How can we get access to the plot of a narrative from data?
- Rising/falling action?
- Invocation of characters?
- Comparison of word frequencies?

# Jack Elliot
## "Vocabulary Decay in Category Romance"

- Category Romance: short novels (200 pages/55k words) with simple plots and major characters and issued on a regular basis.
- Vocabulary decay: the shedding of unique words over the length of the novel.
- What motivates his research?



# Elliot: Workflow

- Archive: 181 Harlequin Presents novellas
- Selection: removed outliers (via PCA), leaving 178 novellas
- Preprocessing: spelling normali(s/z)ation, stripped of preamble and postscript, lowercase (?).
- Segmentation: each text cut into 100 sections based on overall length.
- Transformation: Calculate Shannon Entropy of each segment (measure of lexical richness).
- Transformation: Compared with n-grams (bi and trigrams)

**Abstract**: "Writers of a best-selling category romance imprint share a common tendency to decrease their deployment of unique words over the span of their novels—a phenomenon of 'vocabulary decay'. This tendency cannot be found in the novels of Jane Austen, suggesting this drop is not intrinsic to the romance genre itself, and is unlikely to have any true narrative purpose. A study of Charles Dickens shows that vocabulary decay extends beyond the romance genre. Closer examination reveals vocabulary decay is a result of progressive amounts of linguistic chunking—due to author fatigue or a desire to produce a more readable narrative" (321)

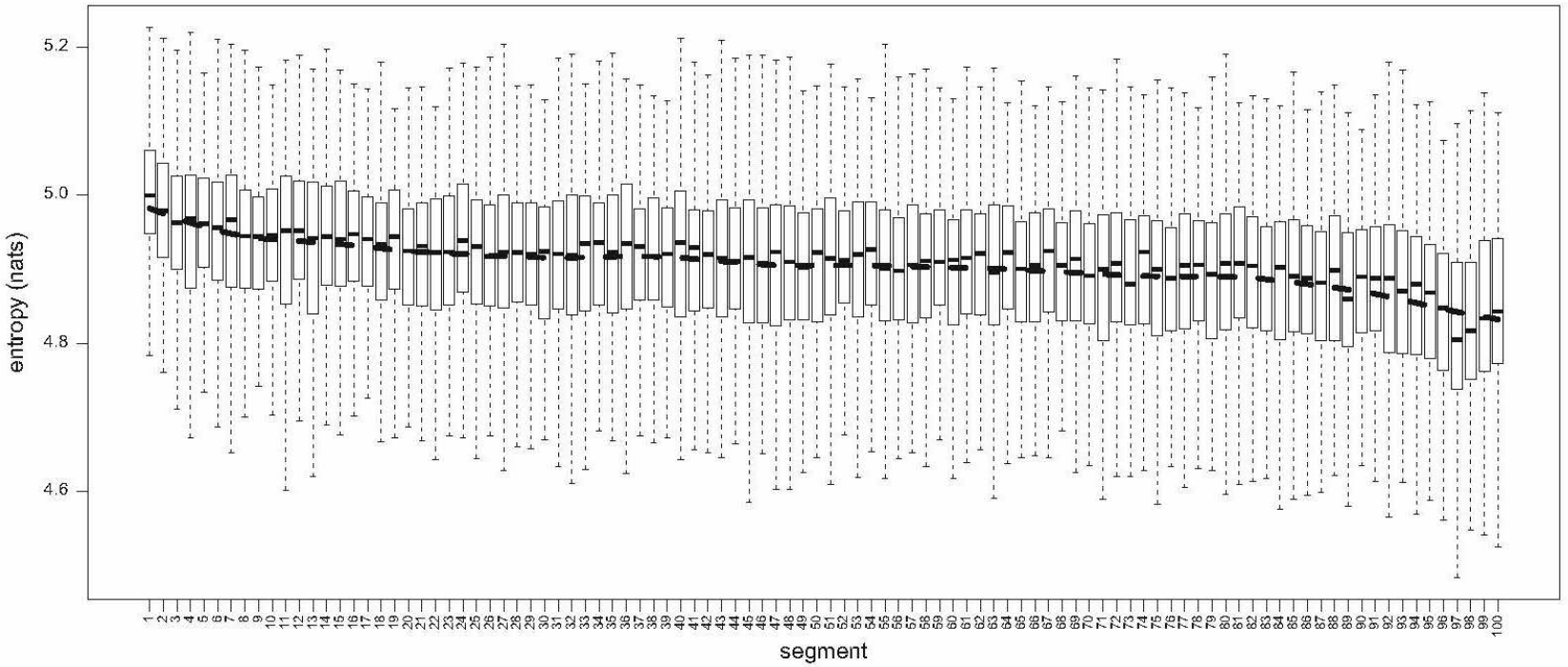



**Fig. 1** Box plot of Shannon Entropy of each segment. Curve has been fitted with Friedman's supersmoother through the mean of each segment. X-axis are segments from 1 (the first) to 100 (the last)

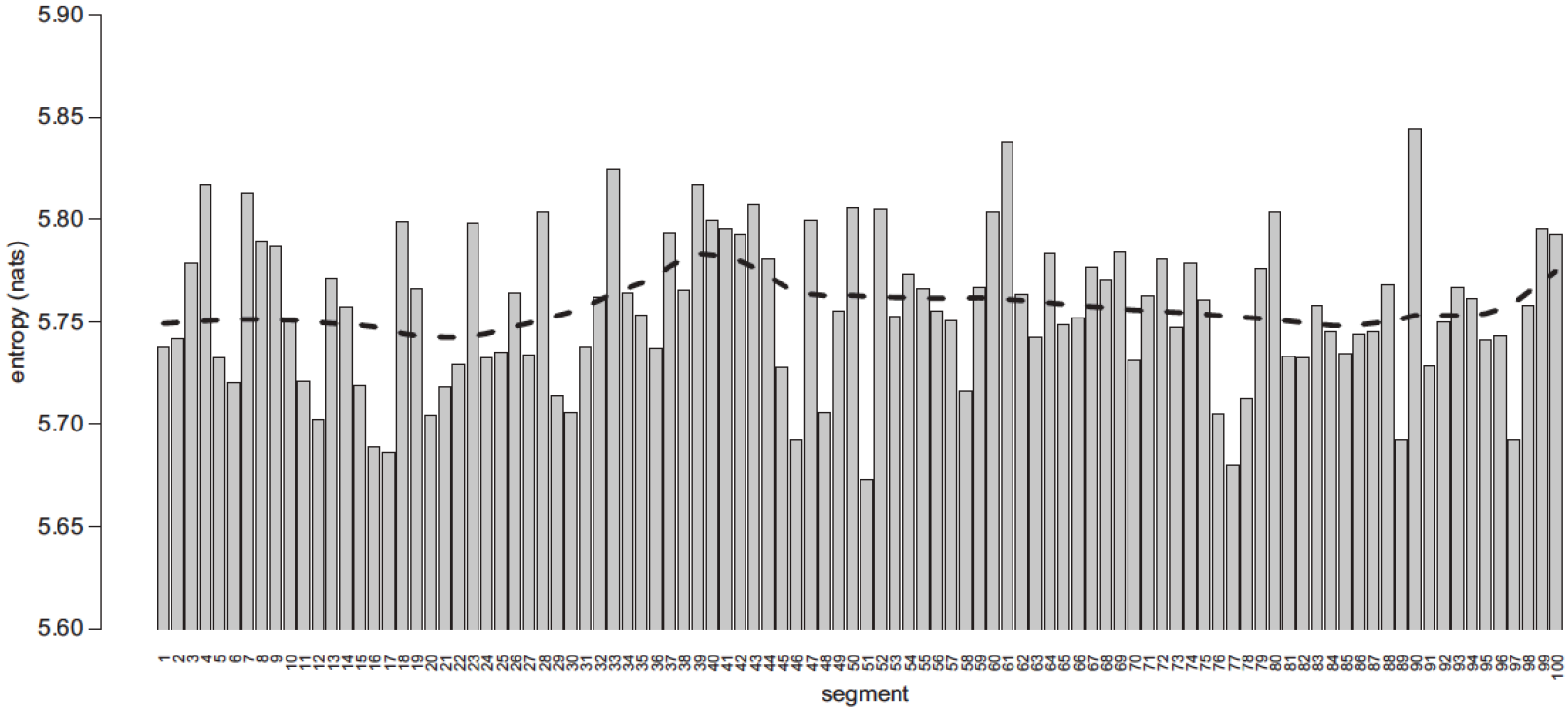
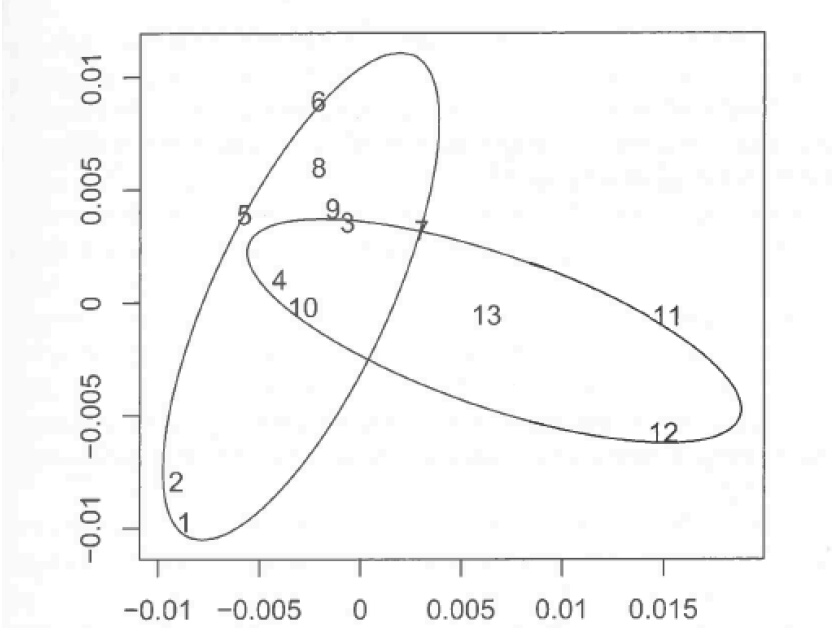



**Fig. 3** Sum of the Shannon Entropy of each segment across Austen's novels. The data are not normally distributed, so means and deviations are not shown

# Andrew Piper
# "Plot (Lack)"

- Tracking major plot shapes.
- Conversion: the before and after model of plot: see St. Augustine's *Confessions*.




2.1 Semantic relationships between the thirteen books of Augustine's *Confessions*. Ellipses have been drawn around the pre- and post-conversional books.

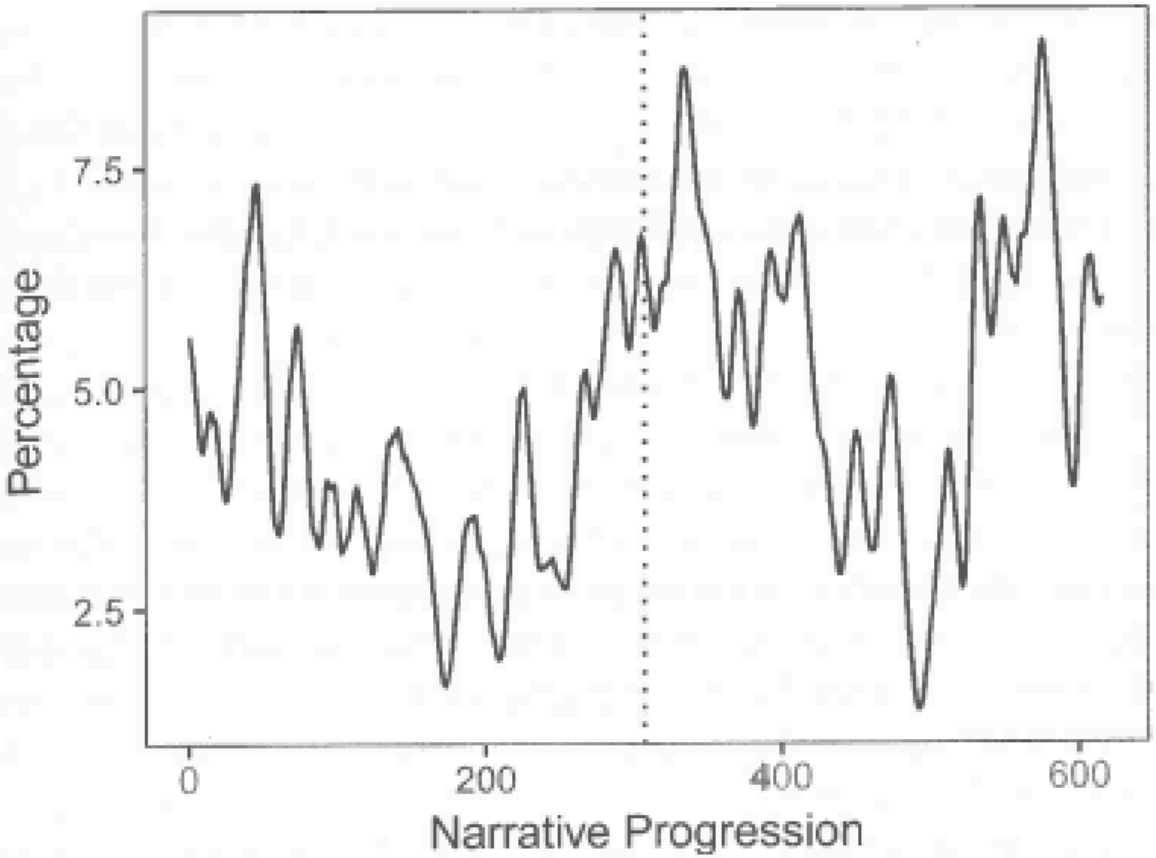
Piper argues that "understanding texts as word distributions gives us an important new way of thinking about plot, the way actions and beliefs are encoded in narrative form. Comparing word distributions allows us to think about the shift or drift of language in a text and the way such transformations signal to readers a change in the text's concerns" (43).

# Piper's Method: Confessions

- Archive: Latin-language edition of Confessions.
- Segmentation: Split the text into volumes/chapters
- Preprocessing: remove stopwords (Latin) and words that appear in fewer than half of all books (as in min_df=.4)
- Transformation: Vectorize
- Statistic: Calculate Euclidean distance between each chapter.
- Reduction of dimensionality: Multi-Dimensional Scaling (MDS) to two-dimensions.
- Visualization: plot chapters in space

```
latin_stopwords = ["a", "ab", "ac", "ad", "adhic ", "aliqui", "aliquis", "an",
"ante", "apud", "at", "atque", "aut", "autem", "cum", "cur", "de", "deinde",
"dum", "ego", "enim", "ergo", "es", "est", "et", "etiam", "etsi", "ex", "fio",
"haud", "hic", "iam", "idem", "igitur", "ille", "in", "infra", "inter", "interim",
"ipse", "is", "ita", "magis", "modo", "mox", "nam", "ne", "nec", "necque",
"neque", "nisi", "non", "nos", "o", "ob", "per", "possum", "post", "pro",
"quae", "quam", "quare", "qui", "quia", "quicumque", "quidem", "quilibet",
"quis", "quisnam", "quisquam", "quisque", "quisquis", "quo", "quoniam",
"sed", "si", "sic", "sive", "sub", "sui", "sum", "super", "suus", "tam", "tamen",
"trans", "tu", "tum", "ubi", "uel", "uero"]
```

# Presence of Proper Names




2.6 Rolling average of proper names in Virginia Woolf's *Mrs. Dalloway*. The dotted line represents the novel's midpoint.