What is text analysis?
What is distant reading?

Christopher Ohge, Martin Steer

Riga Technical University, September 2019

Text Analysis:

A Survey of Principles, Tools, and New Ways of Reading

What is text analysis?

Using computational tools and/or programs to analyse text data of varying sizes by yielding quantitative results and making arguments about historical trends or form, style, content, and context.

This can be broken down further:

Statistical analysis: counting words, calculating word pairings (collocates) and ngrams (groups of two or three words or more co-occurring words), average word and sentence length, mean word usage; organising words that occur too frequently to be studied one by one.

Corpus analysis: searching and pattern recognition across multiple texts.

Linguistic analysis: parts-of-speech tagging, lexical variety and uniqueness, topic modeling, sentiment analysis, stylometry.

Network analysis: mapping connections between metadata and textual data.

Various other visualisations (graphs, maps and trees) for explanatory force.

A variety of calculations (simple character counts, mean word usage, hapax percentages) can help to identify repetition, brevity, and unique word clusters.

Anthony Kenny discusses the “mysterious veneration” that some literary scholars have for single and rare word occurrences, when “the rate of occurrence of a dull common word in a text may be a much more significant feature” (*Computation of Style* [1982], 67–68).

Similarly, **John F. Burrows**, in *Computation into Criticism*, bases his analysis on the 30 most common words in Jane Austen’s novels, with less attention to unique words.

Distant reading and interpretation

Franco Moretti: “Quantitative research provides a type which is ideally independent of interpretations … it provides *data*, not interpretation” (*Graphs, Maps, Tress* [Verso, 2007]).

History: shifting the gaze from extraordinary people and events to everyday facts. What literature can be found in large mass of facts?

“Abstraction is not an end in itself, but a way to widen the domain of the literary historian, and enrich its internal problematic” (Moretti, 2).

Distance is a new kind of knowledge—a model

Burrows (2004) argues that the styles of authors come from common words (i. e., articles and prepositions).

“[T]he real value of studying the common words rests on the fact that they constitute the underlying fabric of a text, a barely visible web that gives shape to whatever is being said … The principal point of interest is neither a single stitch, a single thread, nor even a single color but the overall effect. Such effects are best seen, moreover, when different pieces are put side by side.”

The value of experimentation, and the possibilities for new modes of reading. Burrows:“computer-assisted textual analysis can be of value in many different sorts of literary inquiry, helping to resolve some questions, to carry others forward, and to open entirely new ones.”

David Hoover (2013): “the computer’s greatest strengths are in storing, counting, comparing, sorting, and performing statistical analysis. This makes computer-assisted textual analysis especially appropriate and effective for investigating textual differences and similarities” (see “Textual Analysis” <https://dlsanthology.mla.hcommons.org/textual-analysis/>).

Possible directions of travel:

- Testing a hunch, hypothesis, or thesis about an author, text, passage, genre, or period
- Testing the claims of a critical work
- Investigating how and the extent to which authors differentiate the voices of characters or narrators in a work
- Investigating shifts in style and how they change over time
- Investigating the history of an important word, concept, or group of words or concepts over a long time span
- Studying the effects of genre conventions
- Investigating claim of authorship

Overconfidence and the problem of validation

Adam Hammond suggests that one of the failures of distant reading comes from its lack of discoveries and tendency to over-validate its tools. When unique results are generated, they often cannot be verified, in his estimation.
("The double bind of validation: distant reading and the digital humanities' 'trough of disillusionment.'" *Literature Compass* 14.8 (August 2017): e12402.)

Nan Z. Da: distant reading encourages reductive tautological thinking.
Lacks an intermediary scale of meaningful relations between the macro and the micro; ("The Computational Case against Computational Literary Studies", *Critical Inquiry* 45.3 (Spring 2019): 601-639).

Some principles

1. **Relative frequency**: this comes from John Stuart Mill's **principle of concomitant variation**, which states that if an antecedent circumstance is observed to change proportionally with the occurrence of a phenomenon, it is probably the cause of that phenomenon.

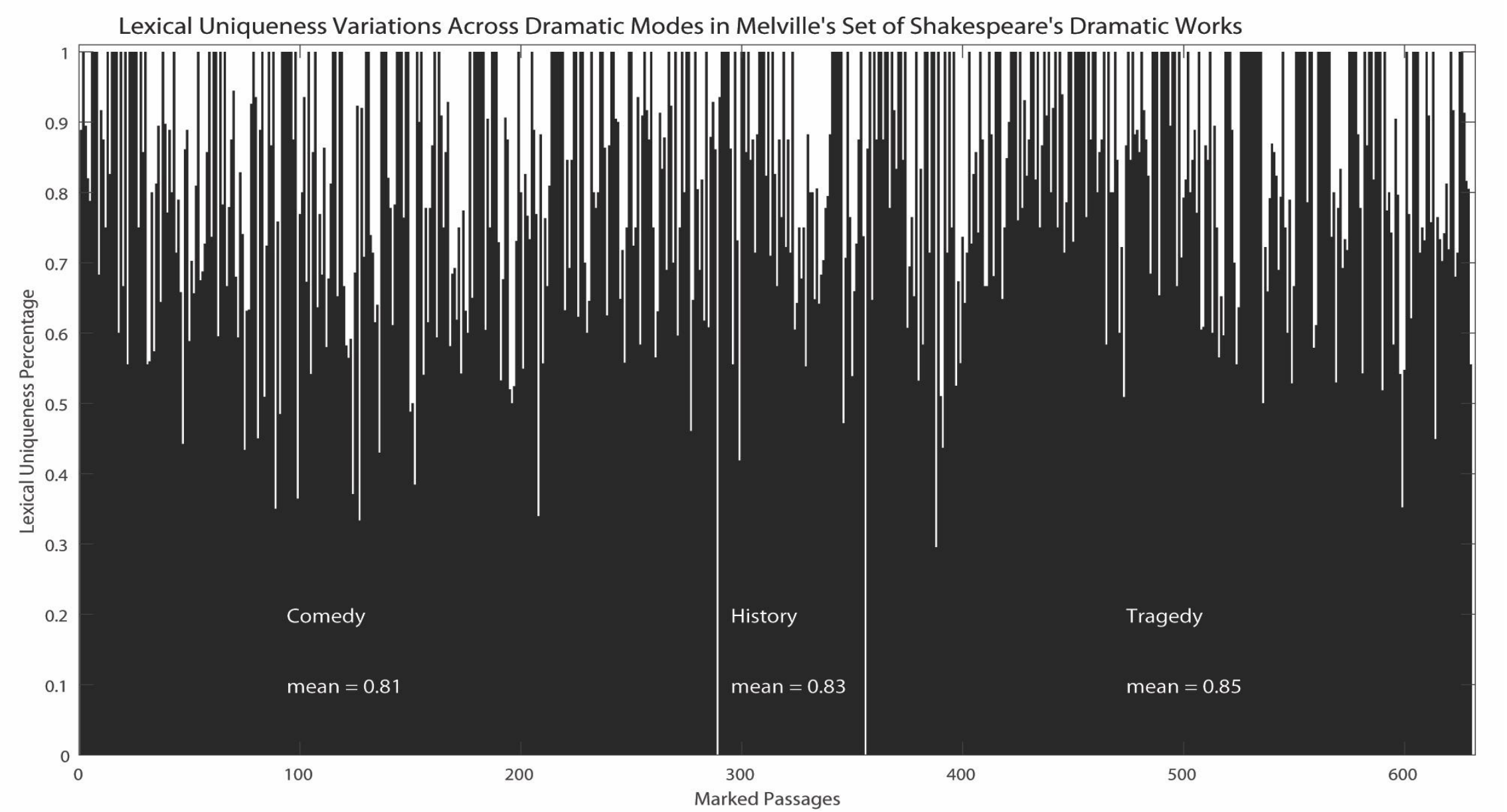
Effects are typically proportional to their causes. In the case of machine learning and word frequencies, this means that texts with similar variables tend to be similar and oftentimes causally connected (that is, authored by the same person).

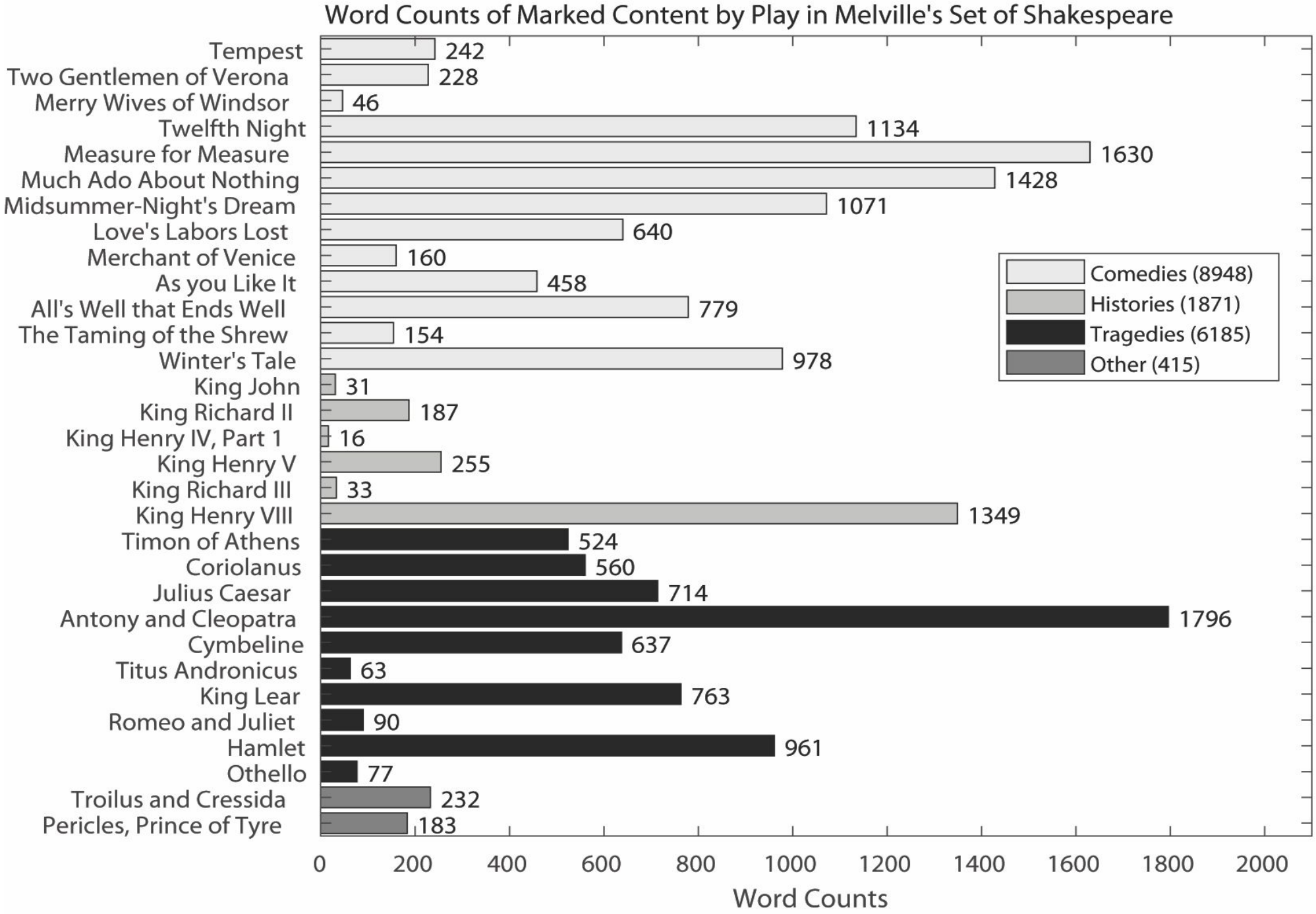
2. **Inductive inference and probability**: creating general hypotheses from specific instances; *yet*, understanding that the results are only probabilistic (i.e., only as good as the quality and comprehensiveness of the data). Justified true belief rather than mere fact.

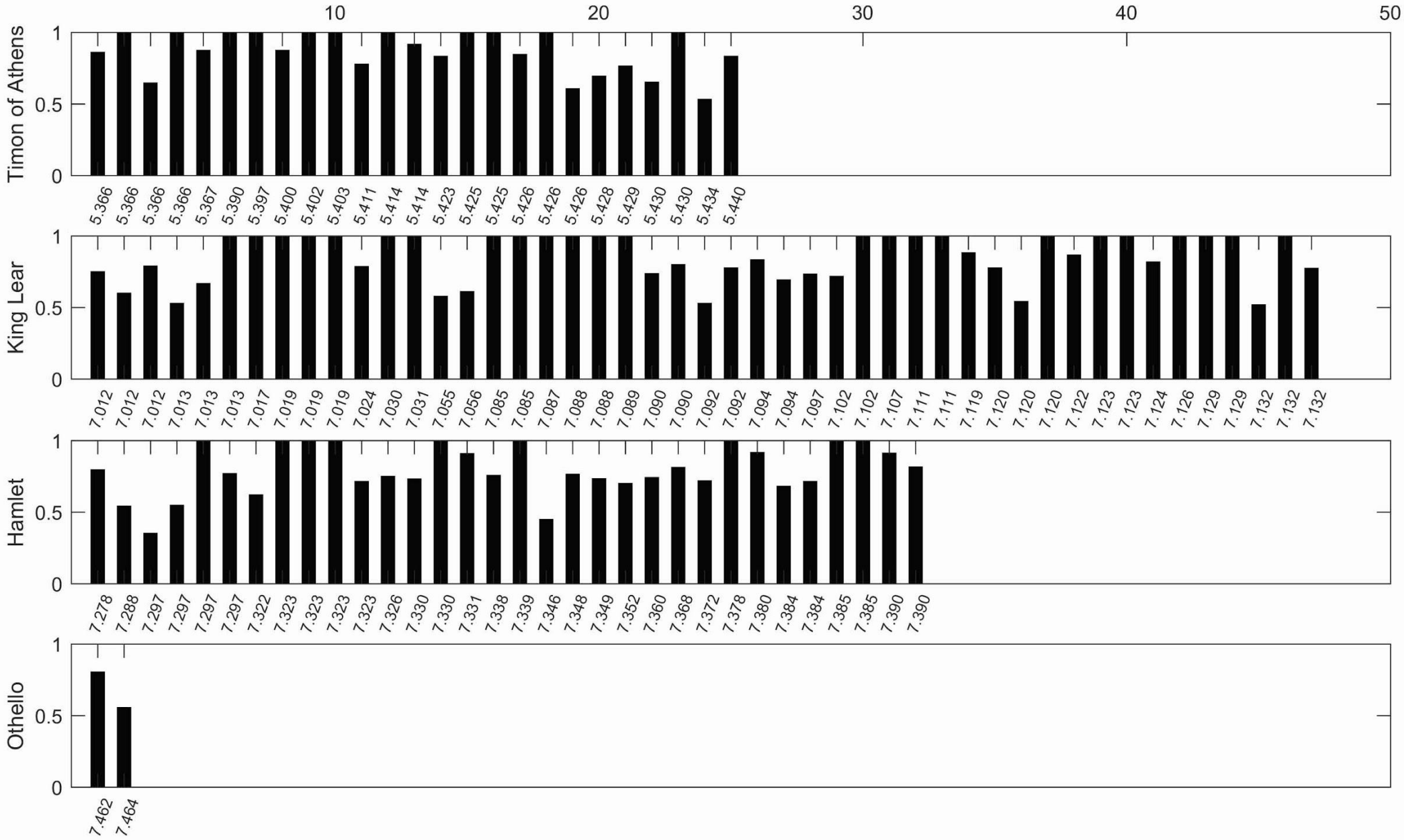
3. **Close reading and critical thinking**. Looking closer at peripheral results, disaffinity and exceptions; letting the research questions guide the exploration; and use skepticism.

4. Distant reading: Making broad claims about historical text data by analysing samples of huge data sets.

Close reading with computers:
Analysing Herman Melville's reading of Shakespeare's plays

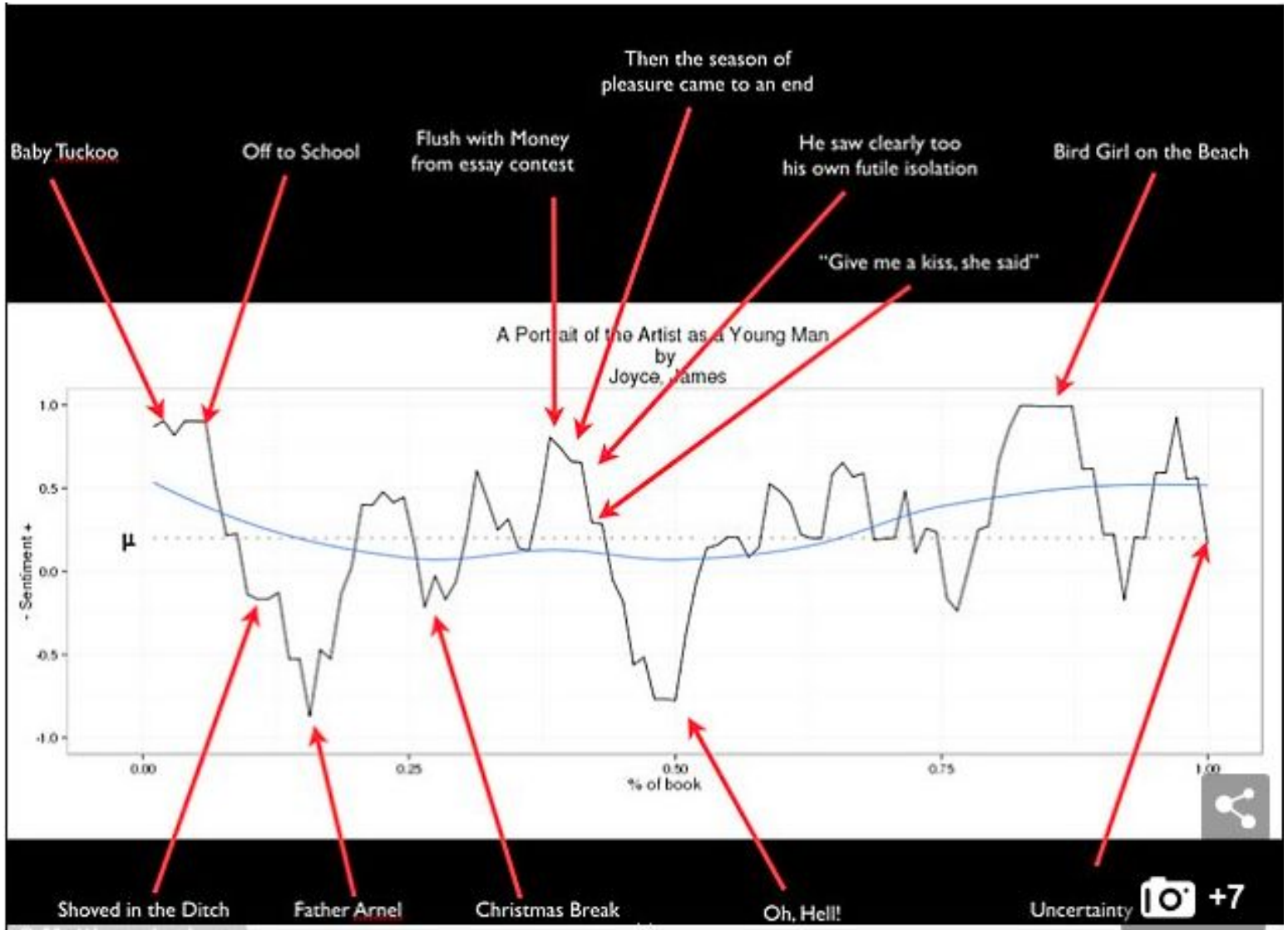



Word Counts of Marked Content by Play in Melville's Set of Shakespeare
Tempest
Two Gentlemen of Verona
Merry Wives of Windsor
Twelfth Night
Measure for Measure
Much Ado About Nothing
Midsummer-Night's Dream
Love's Labors Lost
Merchant of Venice
As you Like It
All's Well that Ends Well
The Taming of the Shrew
Winter's Tale
King John
King Richard II
King Henry IV, Part 1
King Henry V
King Richard III
King Henry VIII
Timon of Athens
Coriolanus
Julius Caesar
Antony and Cleopatra
Cymbeline
Titus Andronicus
King Lear
Romeo and Juliet
Hamlet
Othello
Troilus and Cressida
Pericles, Prince of Tyre
242
228
46
1134
1630
1428
1071
640
160
458
779
154
978
31
187
16
255
33
1349
524
560
714
1796
637
63
763
90
961
77
232
183
Comedies (8948)
Histories (1871)
Tragedies (6185)
Other (415)
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Word Counts

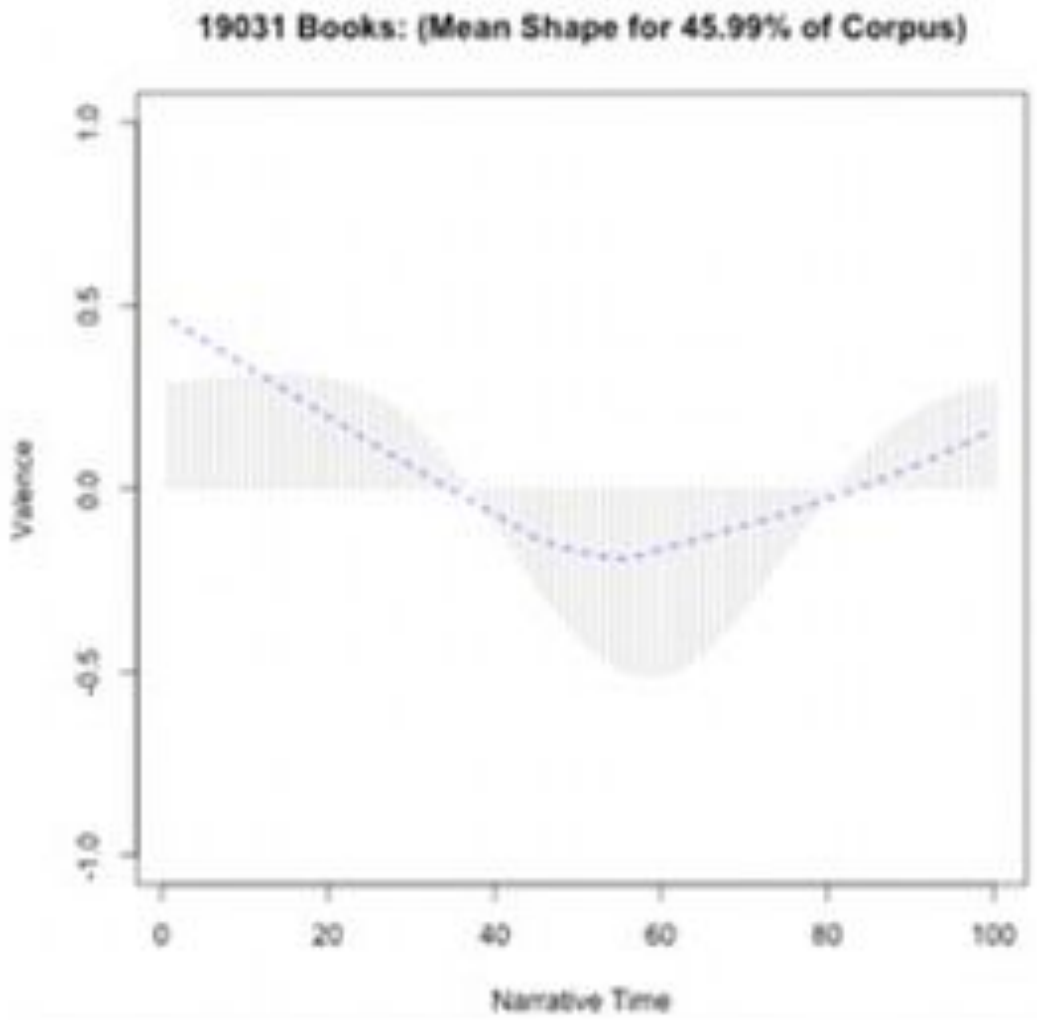

Lexical Uniqueness Values for each Marked Passage in Melville's Marginalia
to Timon of Athens, King Lear, Hamlet, and Othello
10
20
30
40
50
Timon of Athens
King Lear
Hamlet
Othello
1
0.5
0
5.366 5.366 5.366 5.366 5.367 5.390 5.397 5.400 5.402 5.403 5.411 5.414 5.414 5.423 5.425 5.425 5.426 5.426 5.426 5.428 5.429 5.430 5.430 5.434 5.440
7.012 7.012 7.012 7.013 7.013 7.013 7.017 7.019 7.019 7.019 7.024 7.030 7.031 7.055 7.056 7.085 7.085 7.087 7.088 7.088 7.089 7.090 7.090 7.092 7.092 7.094 7.094 7.097 7.102 7.102 7.107 7.111 7.111 7.119 7.120 7.120 7.120 7.122 7.123 7.123 7.124 7.126 7.129 7.129 7.132 7.132 7.132
7.278 7.288 7.297 7.297 7.297 7.297 7.322 7.323 7.323 7.323 7.323 7.326 7.330 7.330 7.331 7.338 7.339 7.346 7.348 7.349 7.352 7.360 7.368 7.372 7.378 7.380 7.384 7.384 7.385 7.385 7.390 7.390
7.462 7.464

Distant Reading:

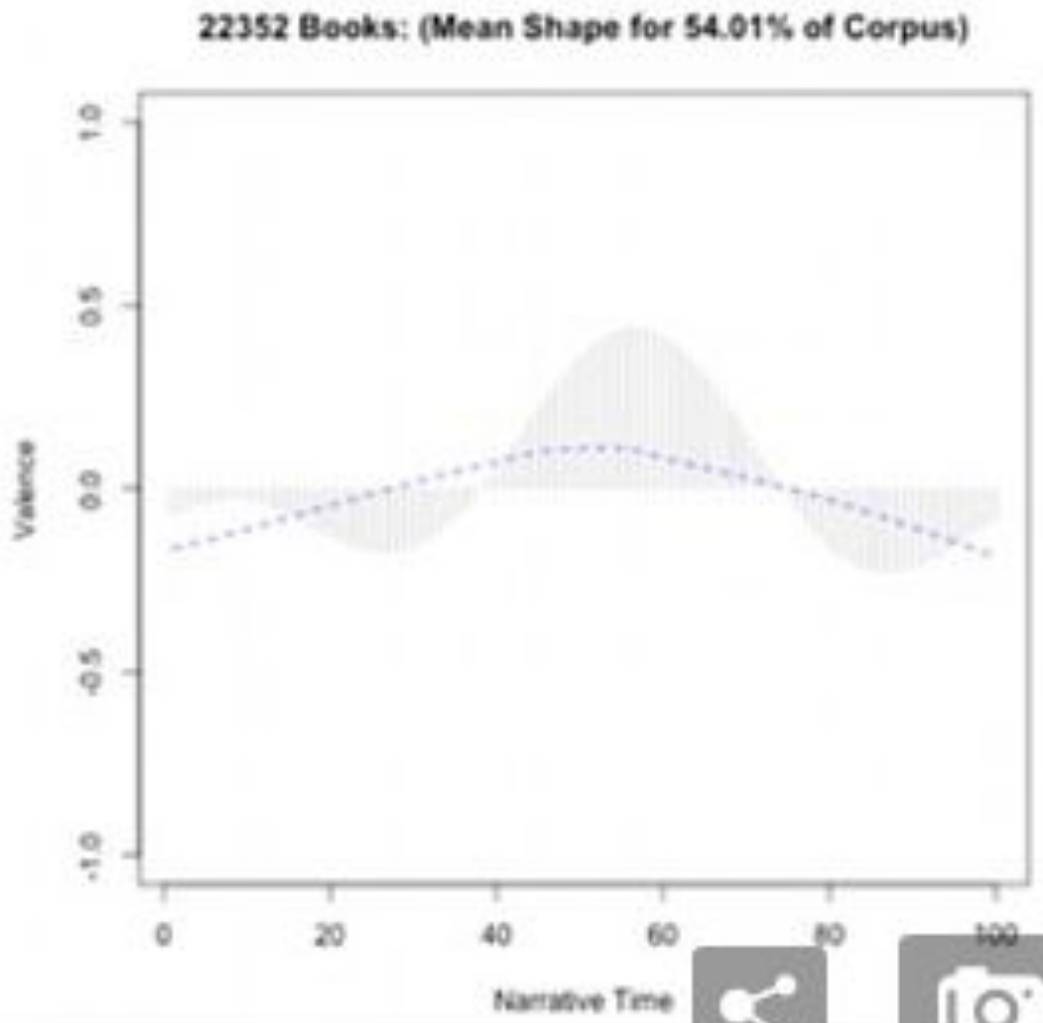
“Professor who analysed 40,000 novels claims there are just SIX possible storylines” (*Daily Mail,* 26 February 2015).

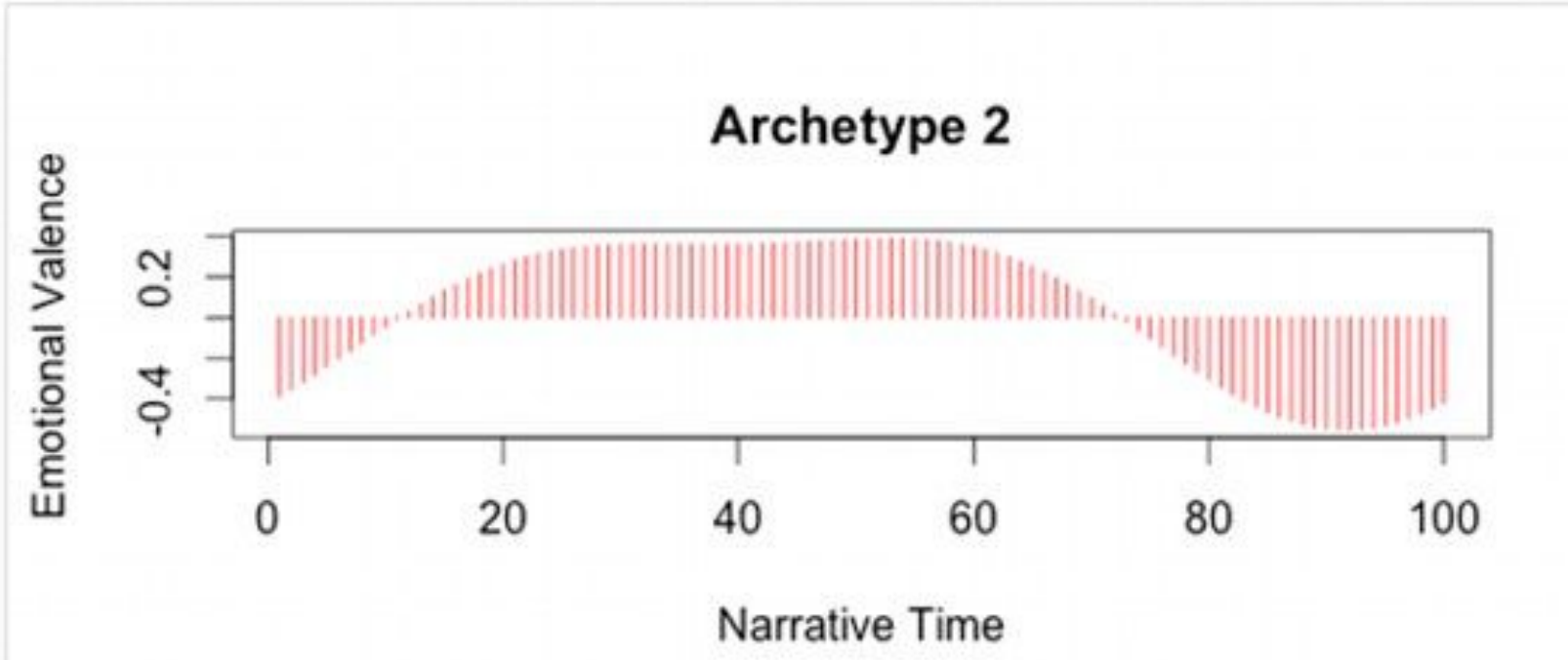


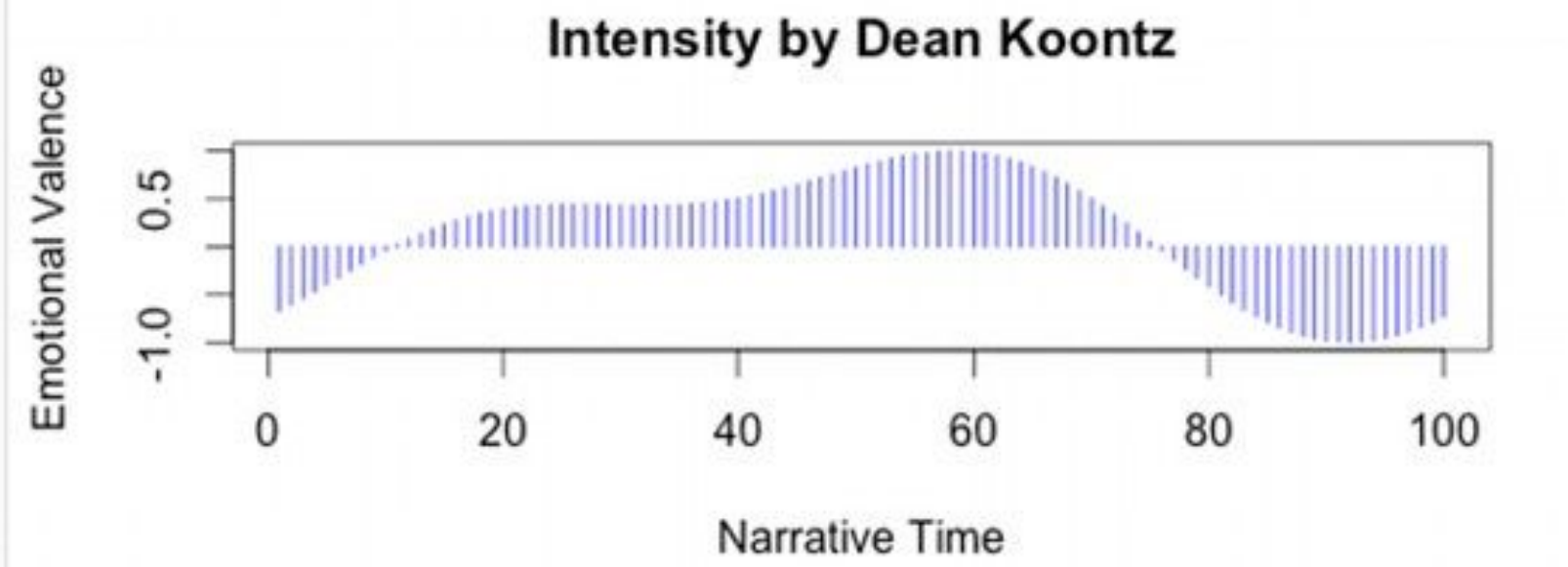
"Man in Hole"



"Man on Hill"




Archetype 2
Emotional Valence
-0.4
0.2
0
20
40
60
80
100
Narrative Time


Intensity by Dean Koontz
Emotional Valence
-1.0
0.5
0
20
40
60
80
100
Narrative Time

Computer-assisted versus computational

Computer-assisted: using out-of-the-box tools to generate results.

Pros: fast results, good interfaces and visualisations.

Cons: lack of control over features and inability to manipulate data; hard to validate and replicate results.

Computational: using programming languages to generate results.

Pros: you have as much control as you can muster; complete customisation.

Cons: requires some knowledge of programming.

Options for performing text analyses

- Quick visualisations: Voyant Tools <https://voyant-tools.org/>
- Corpus creation and analysis: AntConc <http://www.laurenceanthony.net/software/antconc/>
- Language databases: e.g. Historical Thesaurus of English <https://ht.ac.uk/>.
- Text database tools: e.g. Hathi Trust Bookworm and Google nGram searches.
- Programming Language: R or Python (we'll demo R because that is what I know) <https://www.r-project.org/>