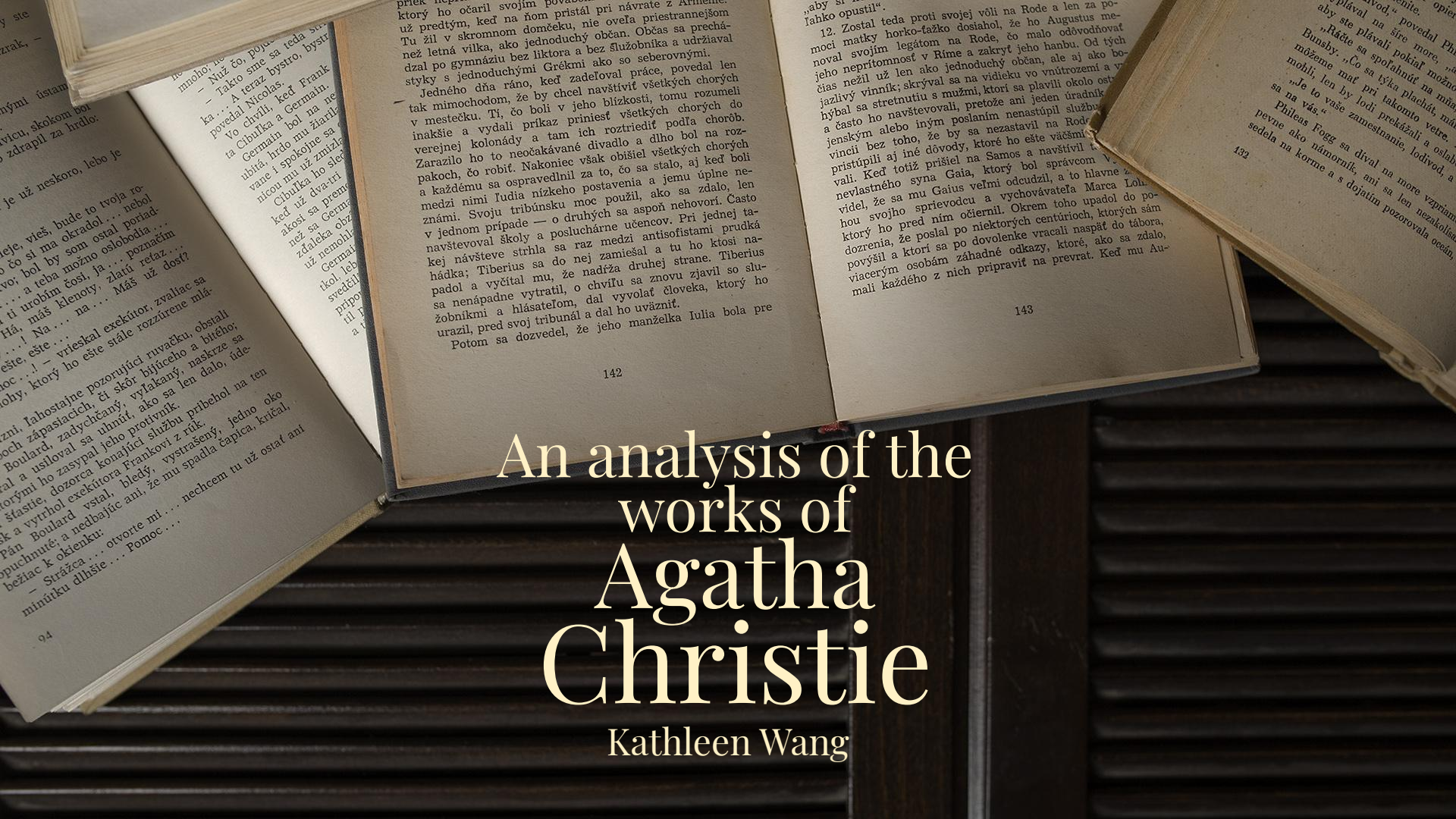


An analysis of the works of Agatha Christie

Kathleen Wang

Background

- 66 detective novels
- 14 short story collections
  - 153 short stories
- 16 plays
- 3 poems
- 2 autobiographies

Background

- Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study

  Ian Lancashire and Graeme Hirst

" … both indefinite words and repetitions occur significantly more often in the language of Alzheimer's patients than in that of healthy people of similar age and level of education."

Problem Statements:

- Can I replicate the findings in the paper?
- Are there other features in the text that are correlated with the age at which the text was written?
- Can I build a model to predict when a piece of Christie’s text was written?

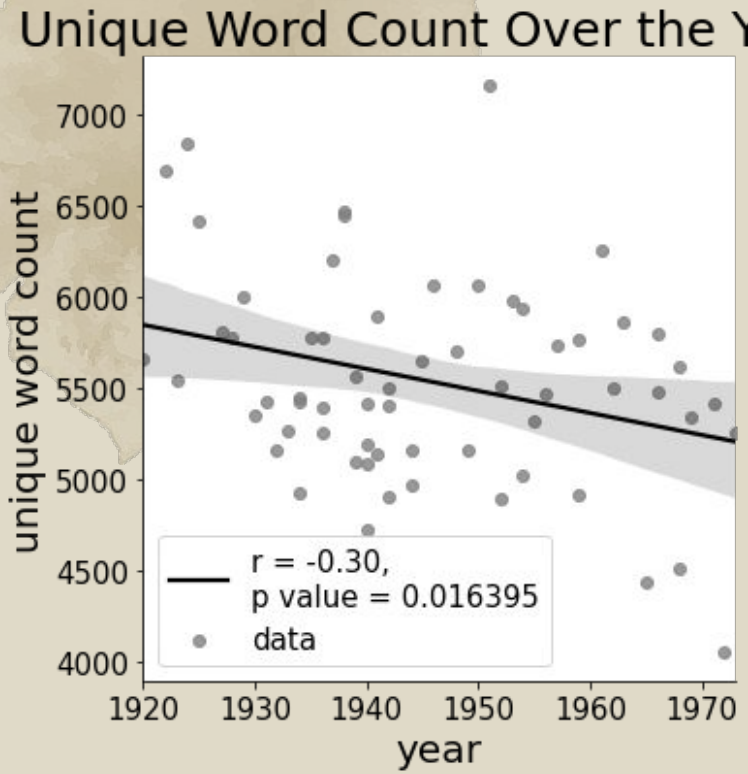
Data

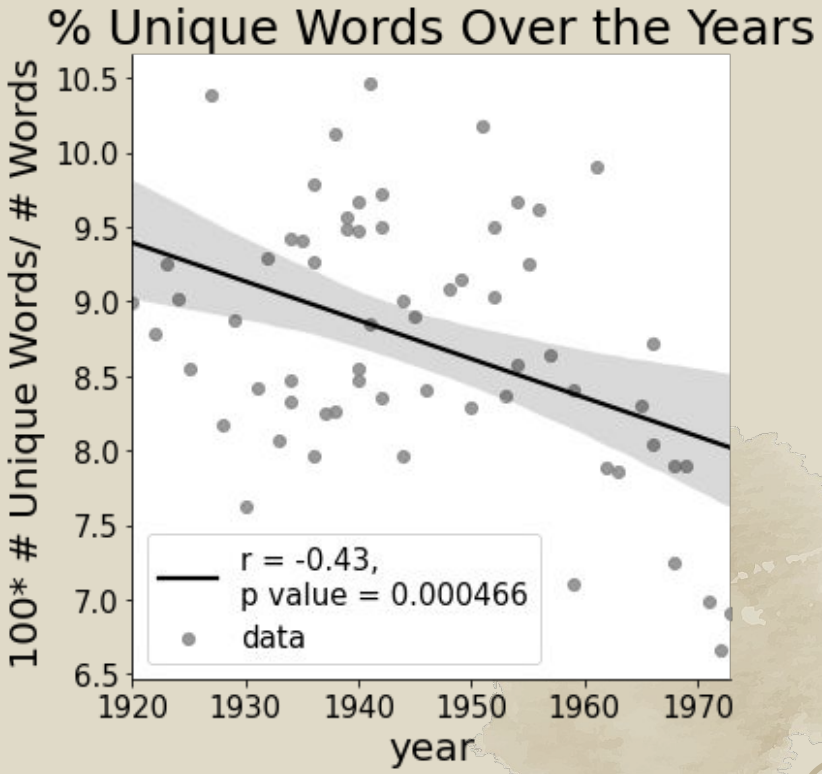
- 60 detective novels from Internet Archive
- Authors used 14 novels
- Cleaning:
  - Extract copyright year
  - Remove forewords/intros
  - Remove punctuation

Outlier

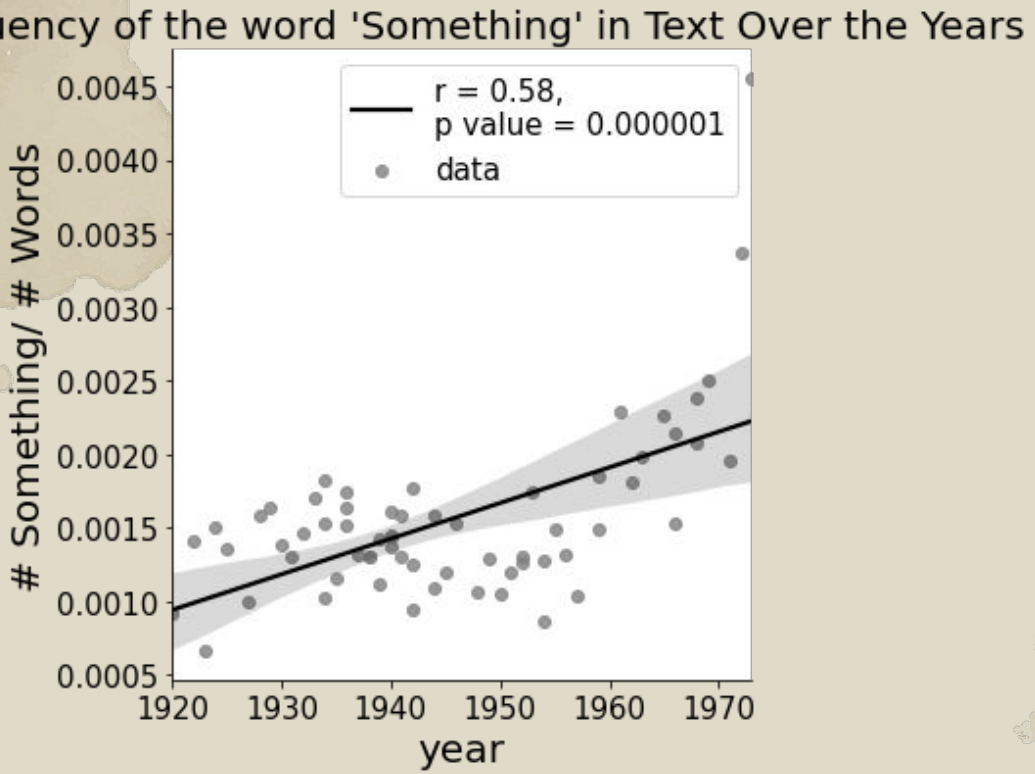
- “Passenger to Frankfurt has the largest vocabulary of all the works we analyzed.”
- “Conceived, written, and researched in her early to mid 70s … draws on books by political thinkers that she requested of her publishers. Much of the vocabulary in Passenger to Frankfurt comes from her reliance on these sources.”

Replicated findings from paper:

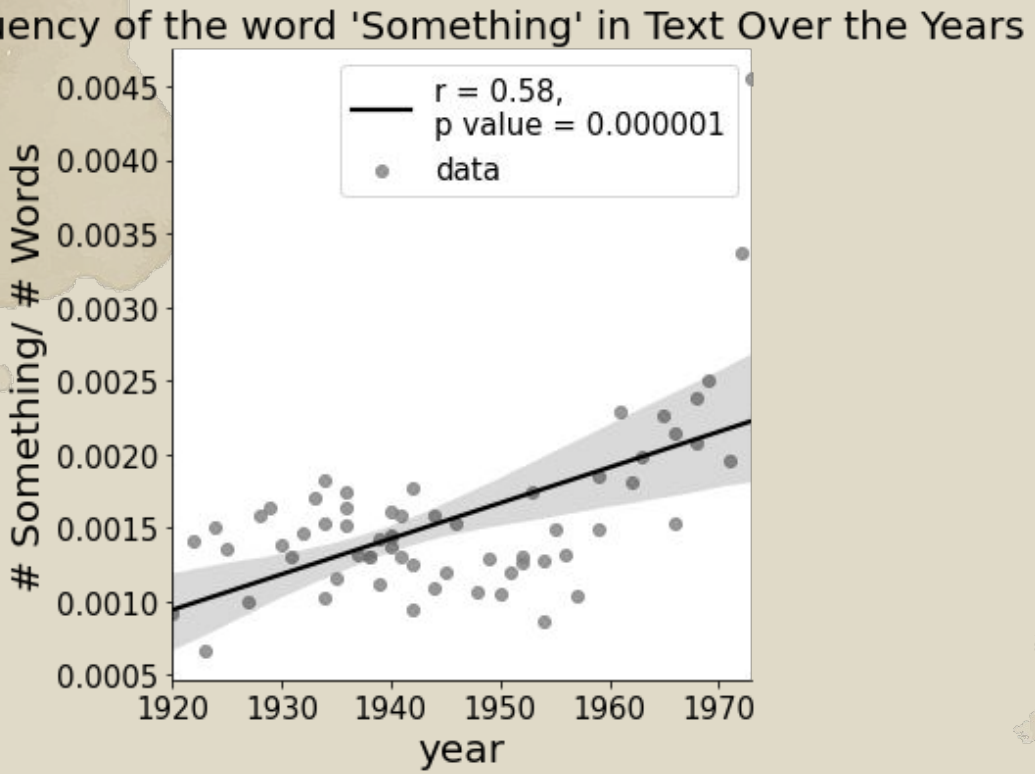




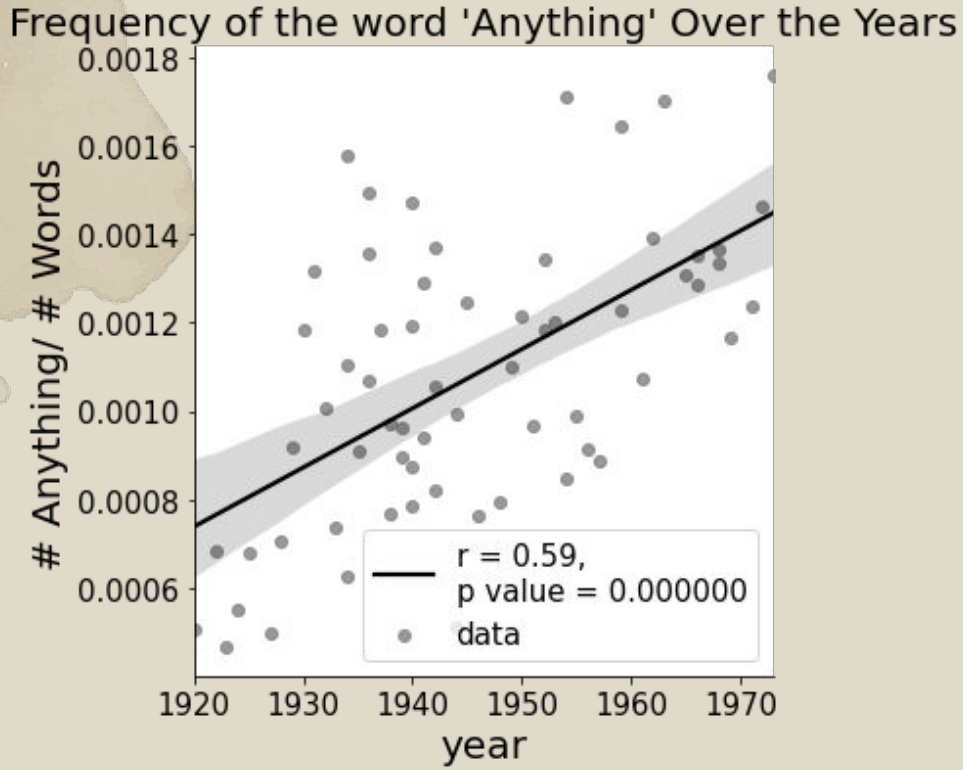
Replicated findings from paper:



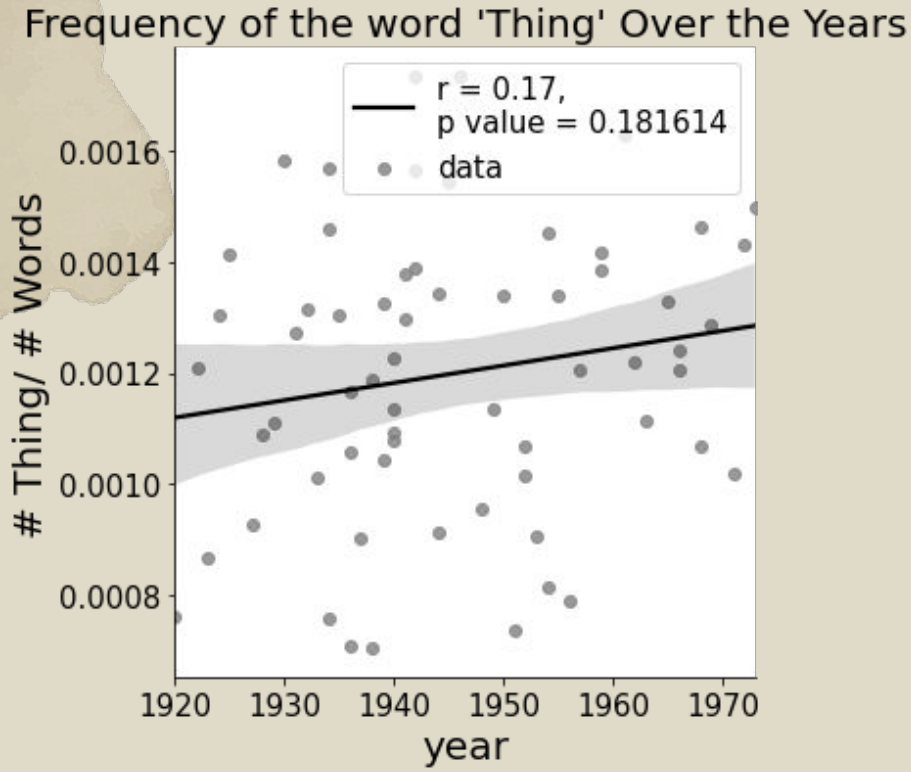
Replicated findings from paper:



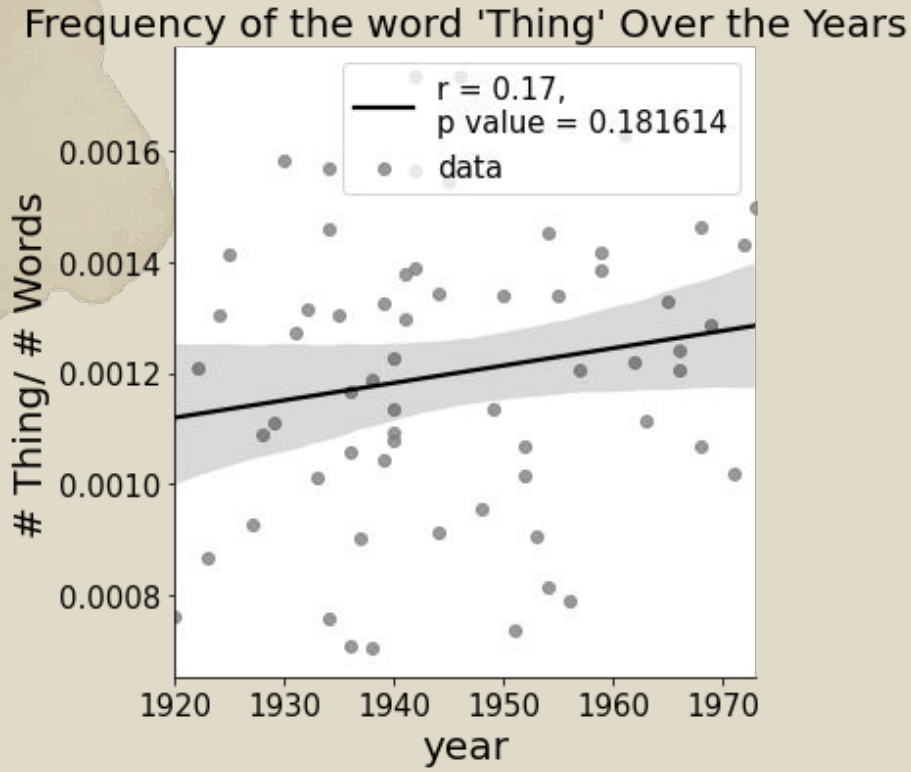
Replicated findings from paper:



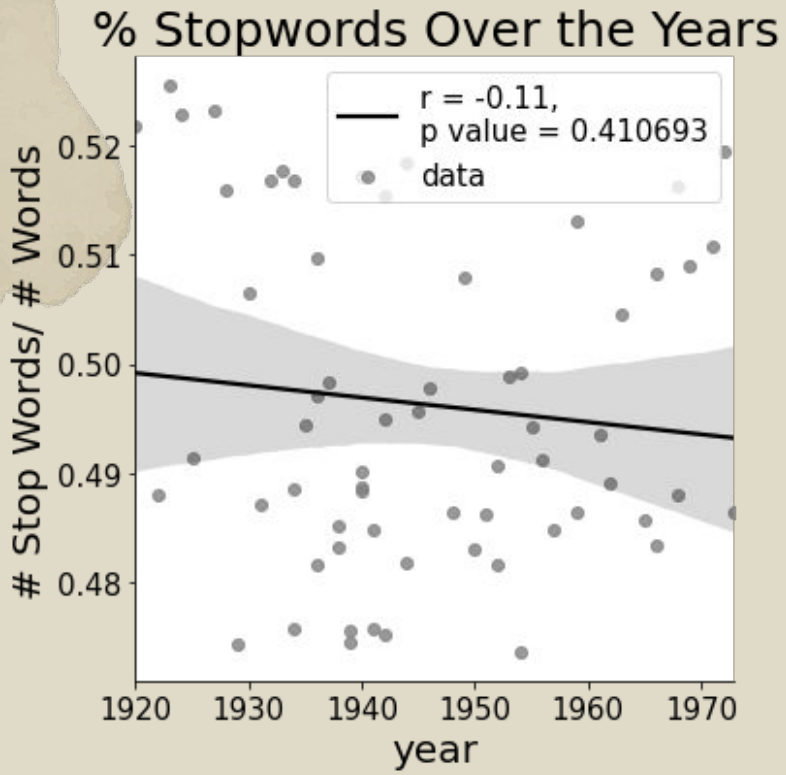
Replicated findings from paper:



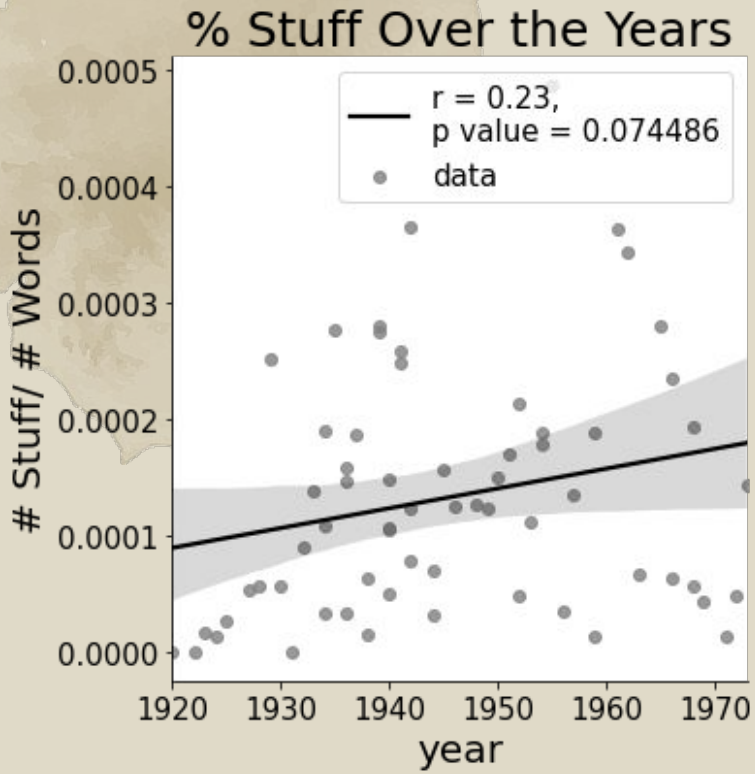
Replicated findings from paper:

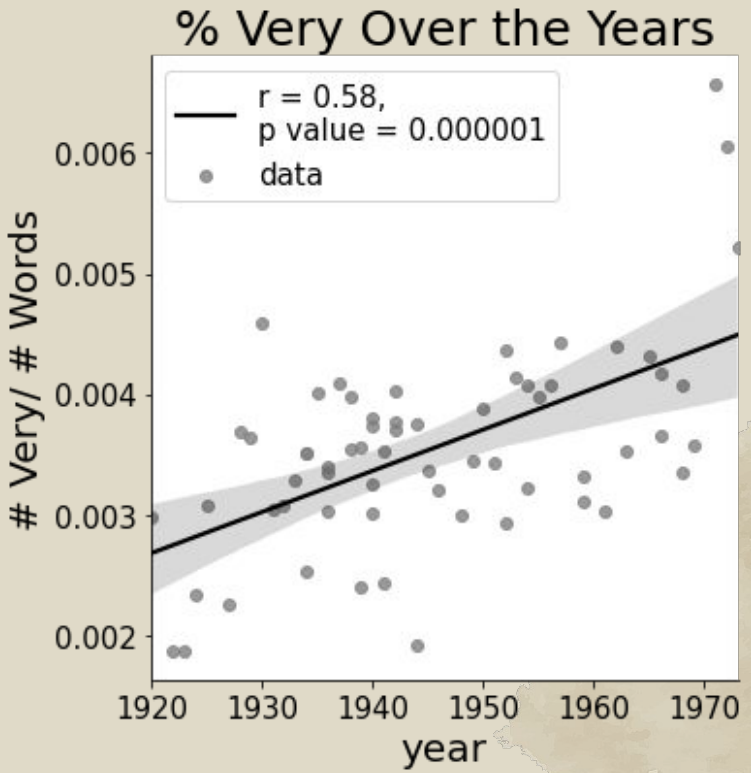


Further Analysis: Stopwords

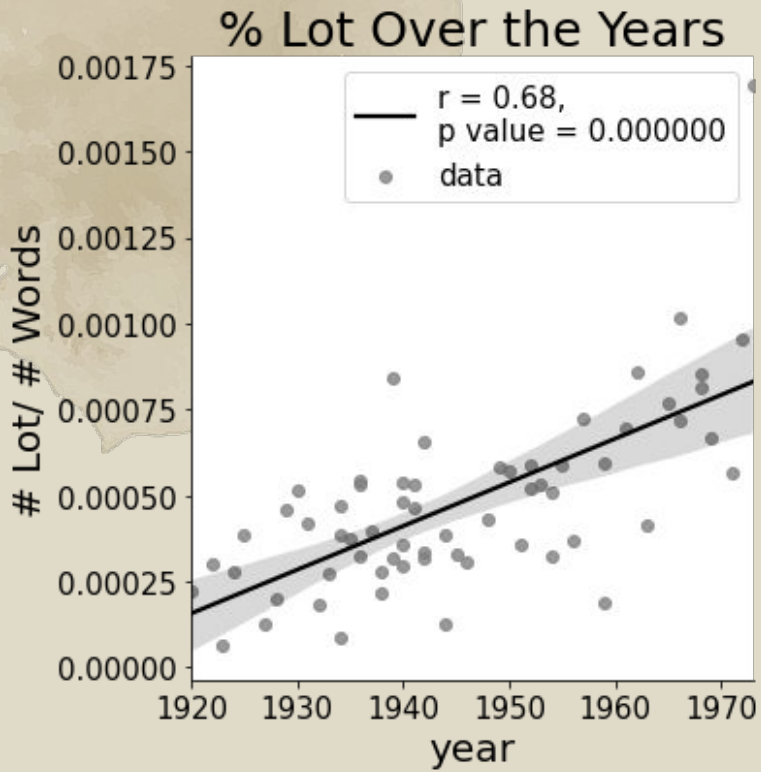


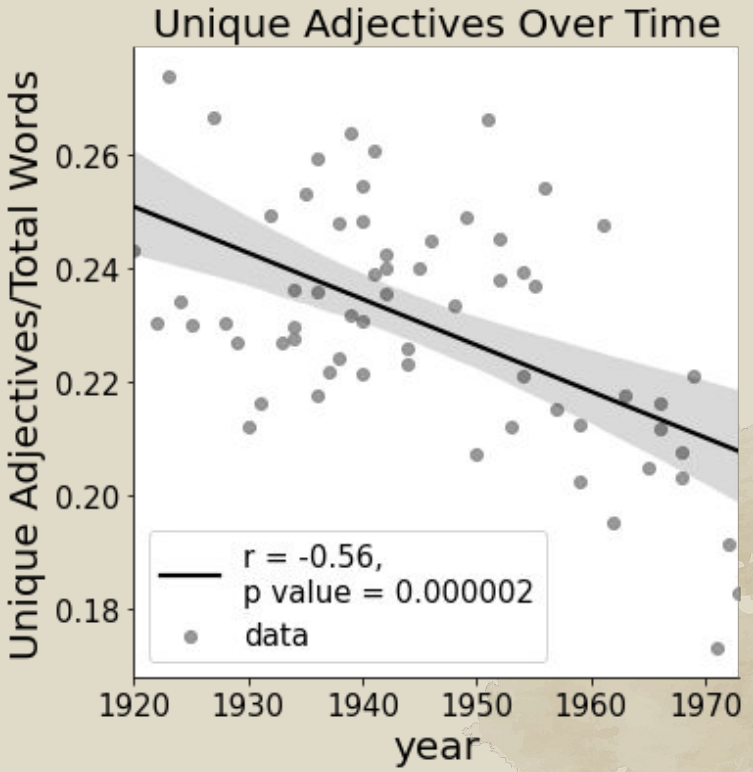
Further Analysis: Other Vague Words





Further Analysis: Other Vague Words





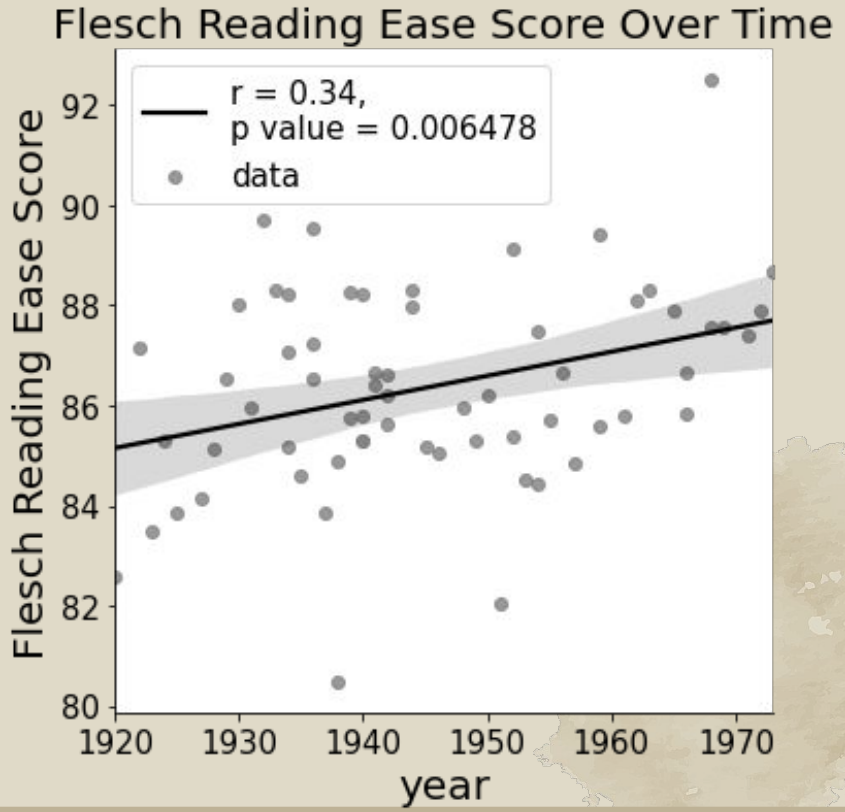
Further analysis: Flesch Reading Score

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

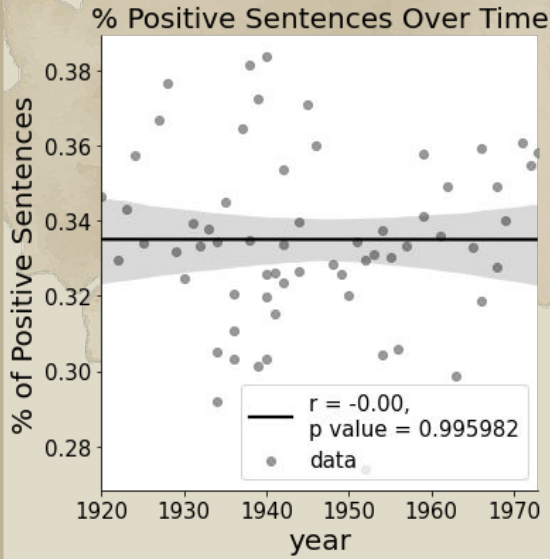
Score	School level (US)	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–10.0	College graduate	Very difficult to read. Best understood by university graduates.
10.0–0.0	Professional	Extremely difficult to read. Best understood by university graduates.

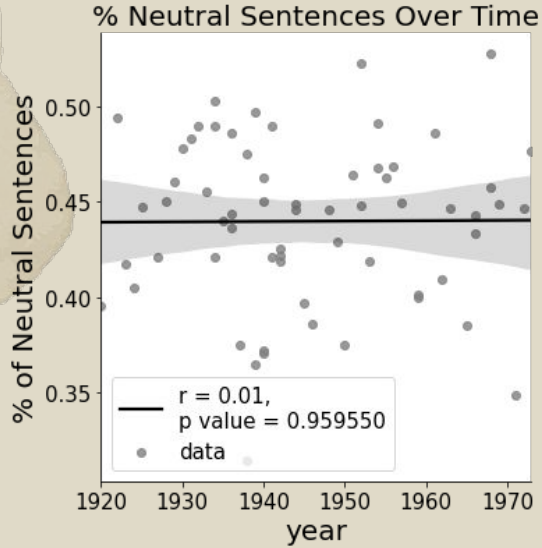
Further analysis: Flesch Reading Score

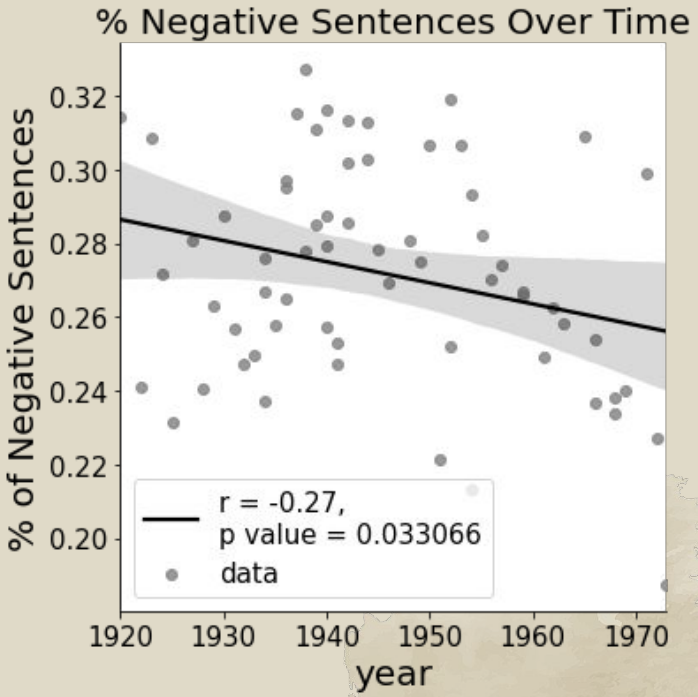
Score	School level (US)
100.00–90.00	5th grade
90.0–80.0	6th grade
80.0–70.0	7th grade
70.0–60.0	8th & 9th grade
60.0–50.0	10th to 12th grade
50.0–30.0	College
30.0–10.0	College graduate
10.0–0.0	Professional



Further analysis: Sentiment Analysis







Modeling:

Model:
Lasso Regression
Train Score: 0.846
Test Score: 0.727

Baseline:
(predicting the average year: 1941)
Train Score: 0.0
Test Score: 0.0

predictions	actual	difference
1976	1973	3
1936	1936	0
1943	1956	13
1945	1949	4
1929	1935	6
1936	1931	5
1930	1927	3
1925	1922	3
1932	1932	0
1934	1941	7
1943	1961	18
1946	1940	6
1944	1955	11
1943	1954	11
1932	1944	12
1952	1963	11

Conclusions:

- There is a decline in Christie's vocabulary recall
- Model would be better with journal entries if you wanted to monitor cognitive decline