SOC 446W Presentation 1

Sophia Hermann, Pavel Ramirez Cammarata, and Carol Zhou

Table of Contents

Harry Potter and the Order of the Phoenix

What is the corpus about?

- The fifth book in the Harry Potter series by J.K Rowling, published in 2003.
- Follows Harry Potter's fifth year at Hogwarts as he faces the growing threat of Lord Voldemort, the oppressive rule of Dolores Umbridge, and struggles with the truth about his connection to the Dark Lord.
- This novel introduces Dumbledore's Army, explores themes of resistance, authority, and adolescence.

Corpus:

- It has 38 chapters, making it the longest book in the series at 257,045 words (Pottermore, 2018)

Intended age range:

- Primarily written for middle-grade and young adult readers (ages 9-18) (Publishers Weekly, 2003)
- Some sources suggest it skews older due to its darker themes and complex narrative (The Guardian, 2016)

Writing style:

- Uses third-person limited narration, mostly from Harry's perspective, allowing readers to experience his emotions and thoughts directly (Mendelsohn, 2018)
- Mixes descriptive prose, internal monologue, and dialogue-heavy scenes, making the story immersive and character-driven (Tiffin & Dowling, 2010).
- Darker tone compared to earlier books, reflecting Harry's emotional struggles and the series' shift toward more mature themes (Granger, 2008).

The Scarlet Letter

What is the corpus about?

- A novel by Nathaniel Hawthorne, first published in 1850.
- Set in Puritan New England during the 17th century, it follows Hester Prynne, a woman forced to wear a scarlet "A" for committing adultery.
- Explores themes of sin, guilt, redemption, and societal judgment.
- Examines moral hypocrisy and the effects of public shaming.

Number of chapters:

- 24 chapters plus an introductory essay, The Custom-House, which provides context for the story (Hawthorne, 1850).

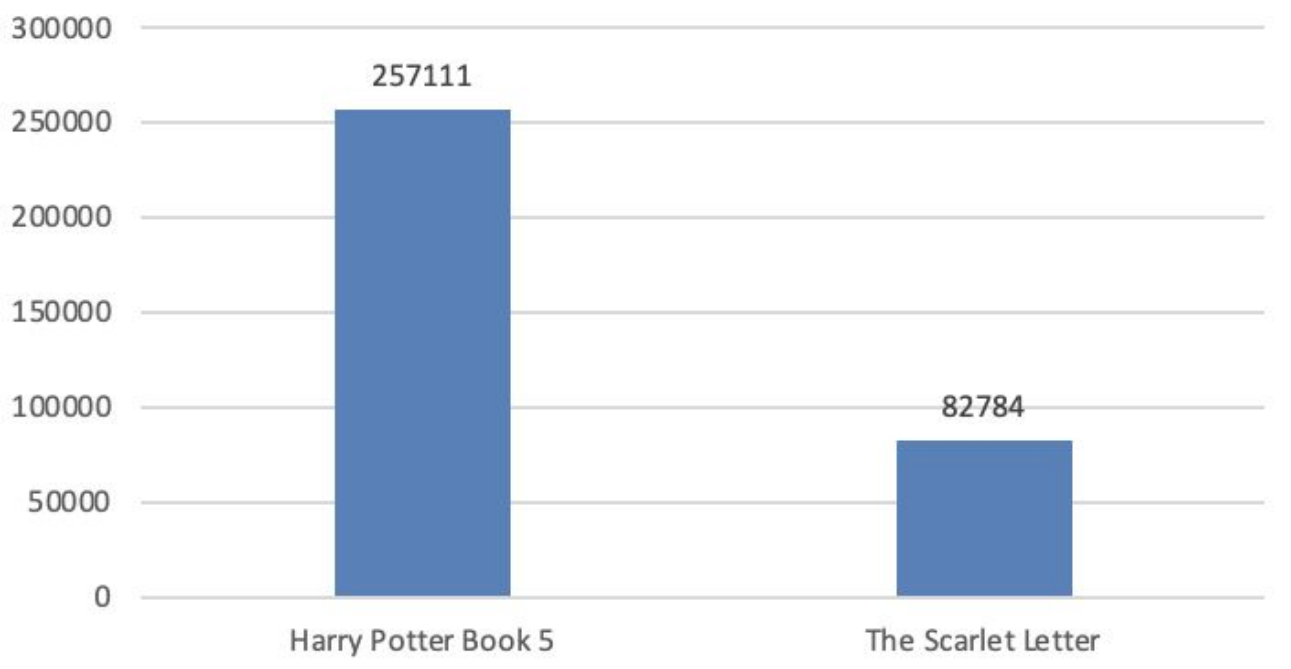
Intended age range:

- Typically assigned to high school and college students (ages 16+) due to its complex themes and advanced language (College Board, 2021).
- Considered part of American literary canon, often included in AP Literature curricula (American Library Association, 2019).

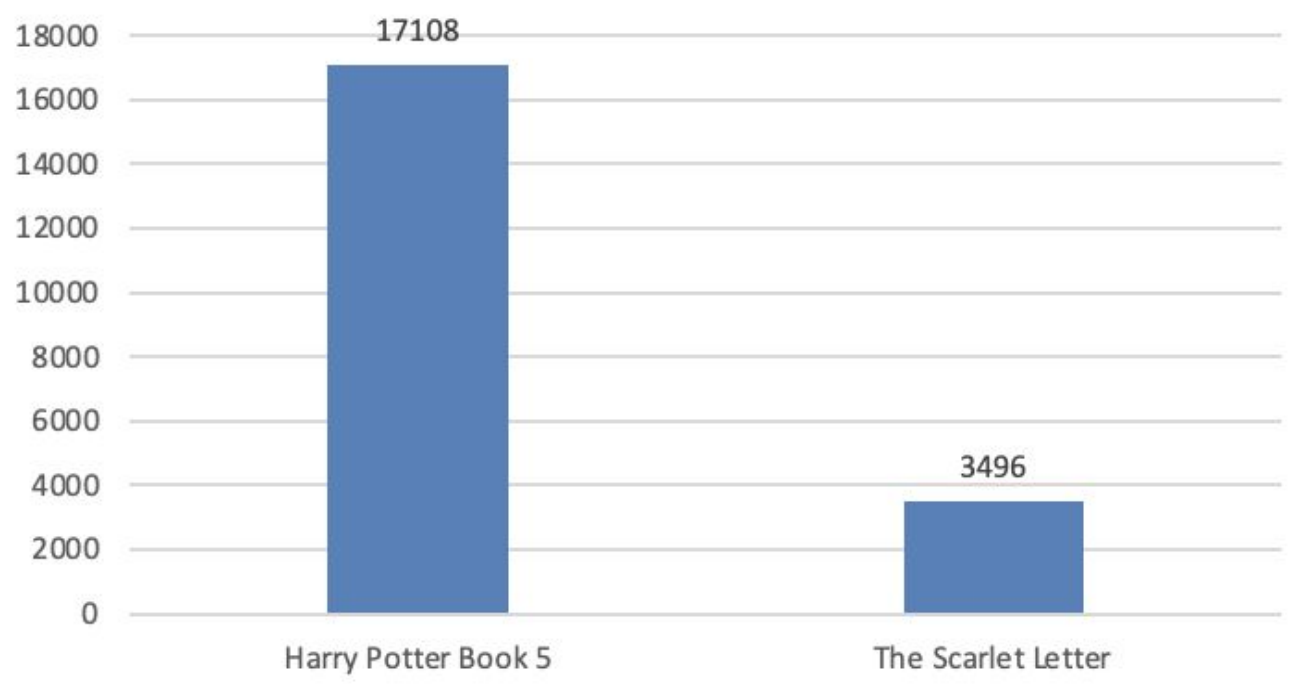
Writing style:

- Formal and highly symbolic, characteristic of Dark Romanticism (Reynolds, 2008).
- Heavy use of allegory, biblical allusions, and intricate sentence structures (Person, 2007).
- Narration is omniscient third-person, offering deep psychological insights into characters (Bercovitch, 1991).

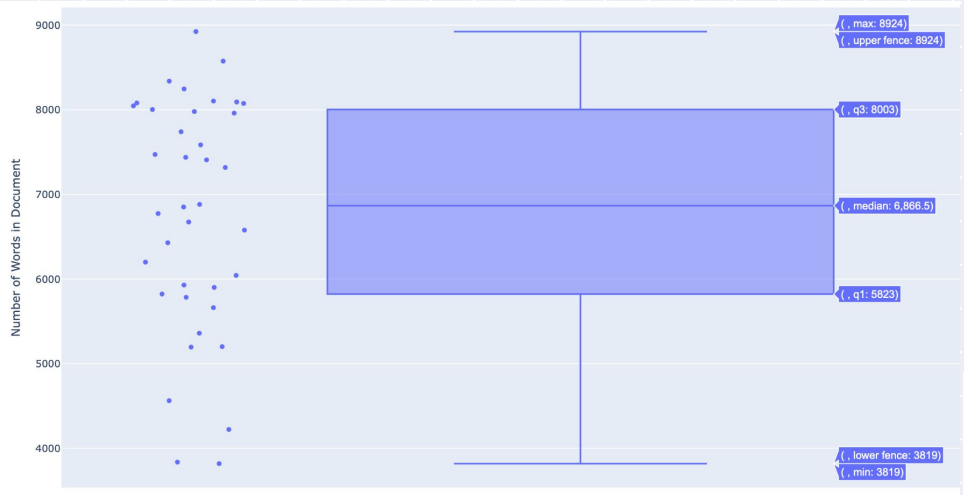
Corpus Statistics — Total Word Count



Corpus Statistics — Total Sentence Count



Corpus Statistics — Word Count per

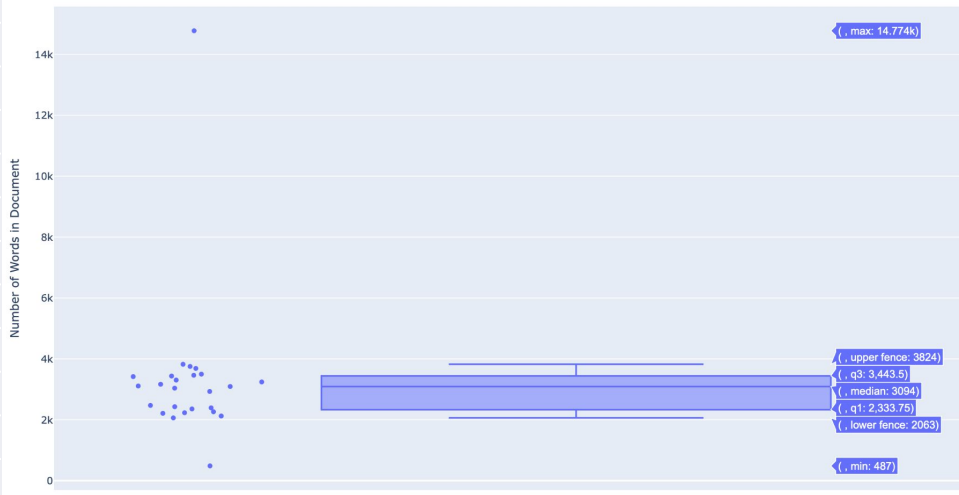


Harry Potter Book 5

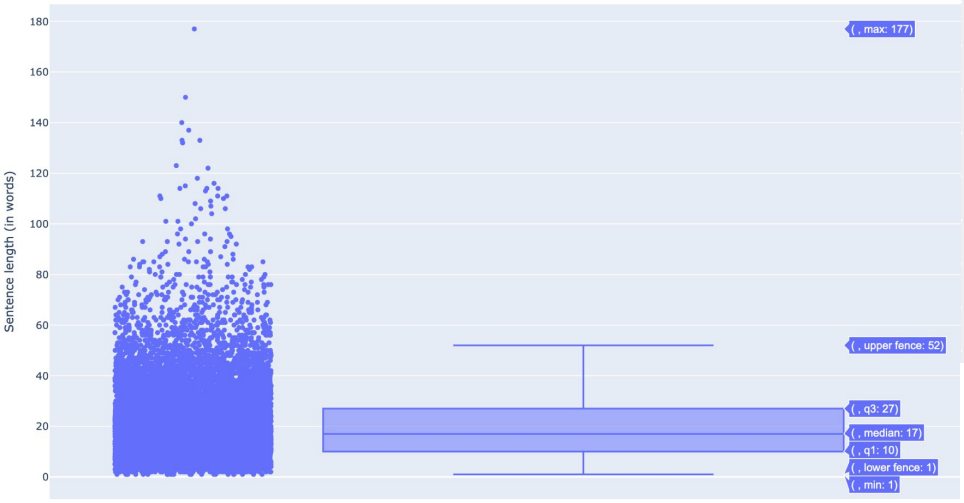
- Max: 8924
- 3rd quartile: 8003
- **Median: 6866.5**
- 1st quartile: 5823
- Min: 3819

The Scarlet Letter

- Max: 14.774k
- 3rd quartile: 3443.5
- **Median: 3094**
- 1st quartile: 2333.75
- Min: 487



Corpus Statistics — Sentence Length

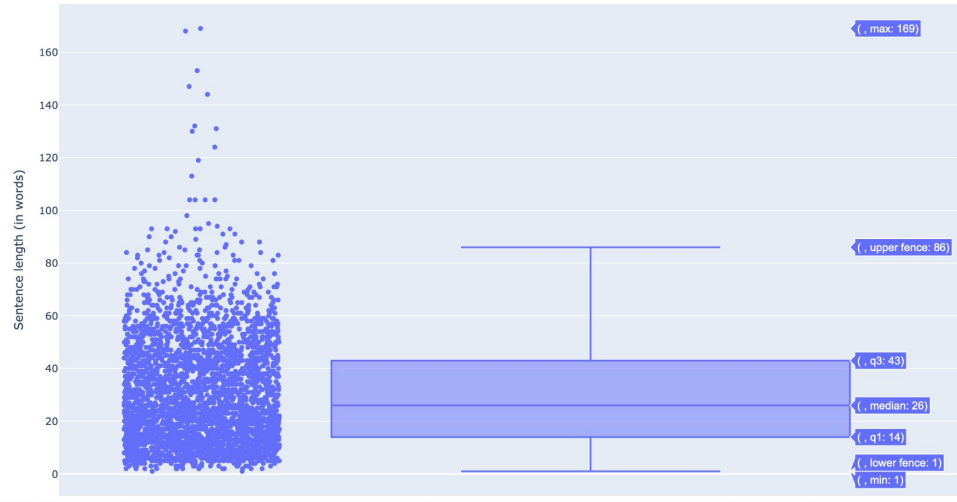


Harry Potter Book 5

- Max: 177
- 3rd quartile: 27
- **Median: 17**
- 1st quartile: 10
- Min: 1

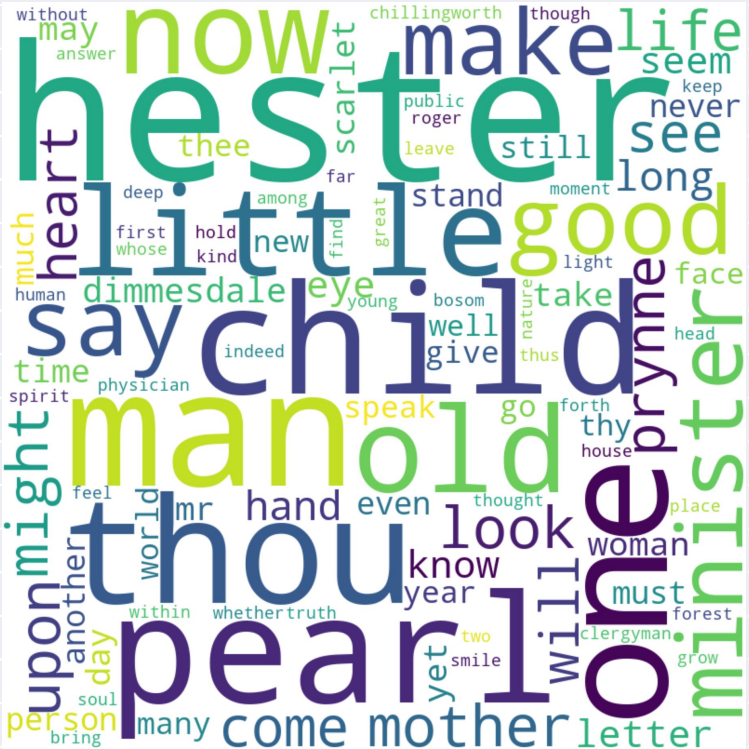
The Scarlet Letter

- Max: 169
- 3rd quartile: 43
- **Median: 26**
- 1st quartile: 14
- Min: 1



Word Clouds

Harry Potter Book 5



The Scarlet Letter

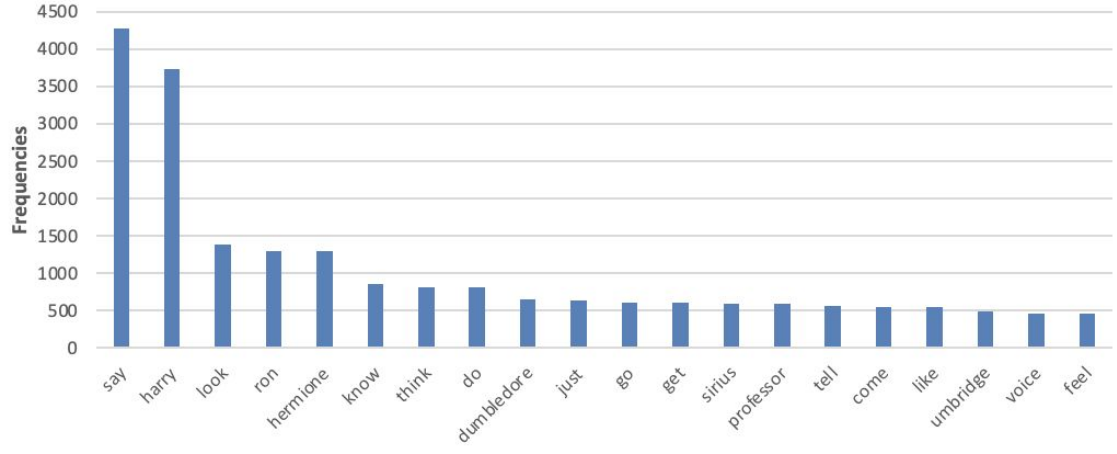
N-grams

“A sequence of N words” [1]

- Unigram consists of a single word (e.g., “machine”)
- Bigram consists of two words (e.g., “machine learning”)
- Trigram consists of three words (e.g., “advanced machine learning”), and so on

Useful to identify differences and similarities in style, tone, and language structure

N-grams — Unigrams

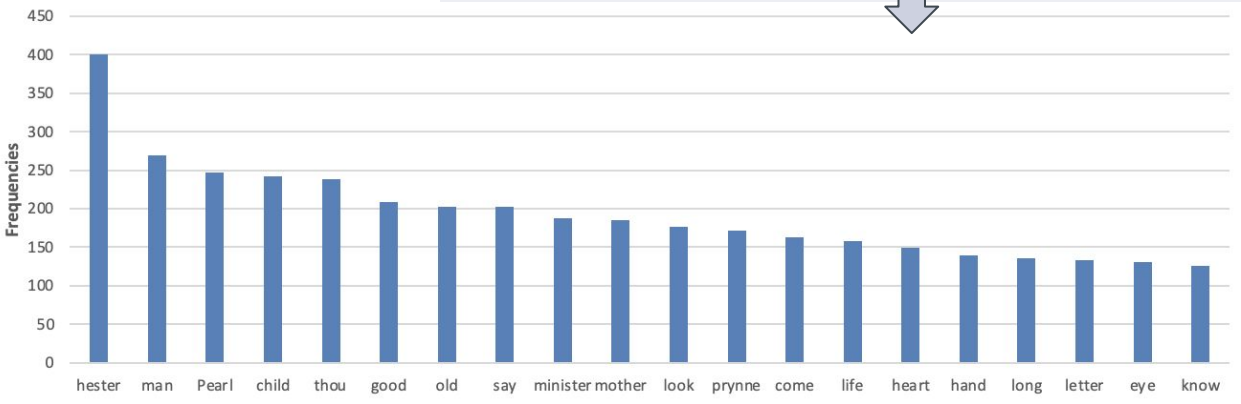


Harry Potter Book 5

- Verbs: “say”, “look”, “come”, and “know”
- Character names: “Harry”, “Ron”, and “Hermione”
- Context-specific: “professor”

The Scarlet Letter

- Verbs: “say”, “look”, “come”, and “know”
- Character names: “Hester”, “Pearl”, and “Dimmesdale”/“minister”
- Context-specific: “mother”, “heart”, and “letter”
- “thou” = “you”



Similarities:

- Focus on dialogue and action
- Character-driven narratives

Differences:

- Context-specific words
- Archaic language

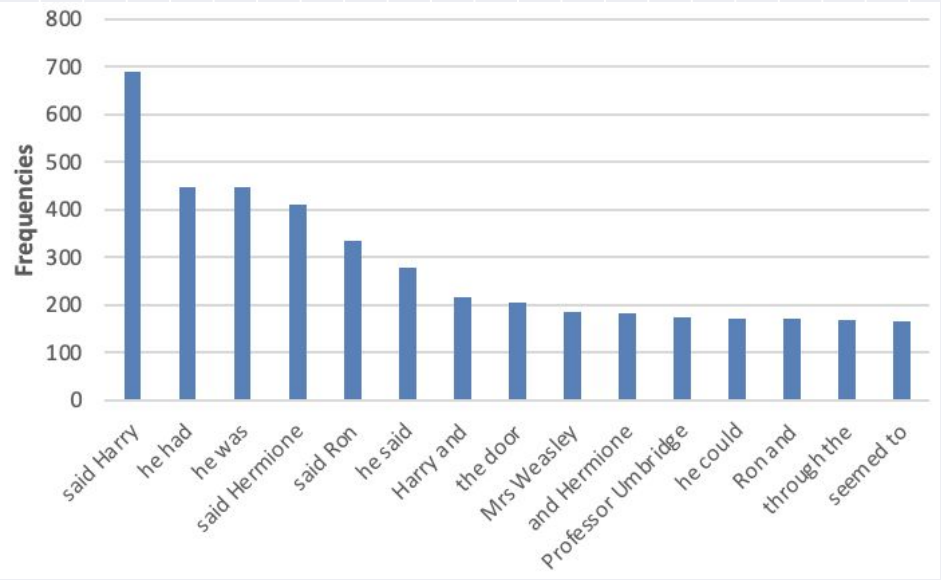
N-grams — Bigrams

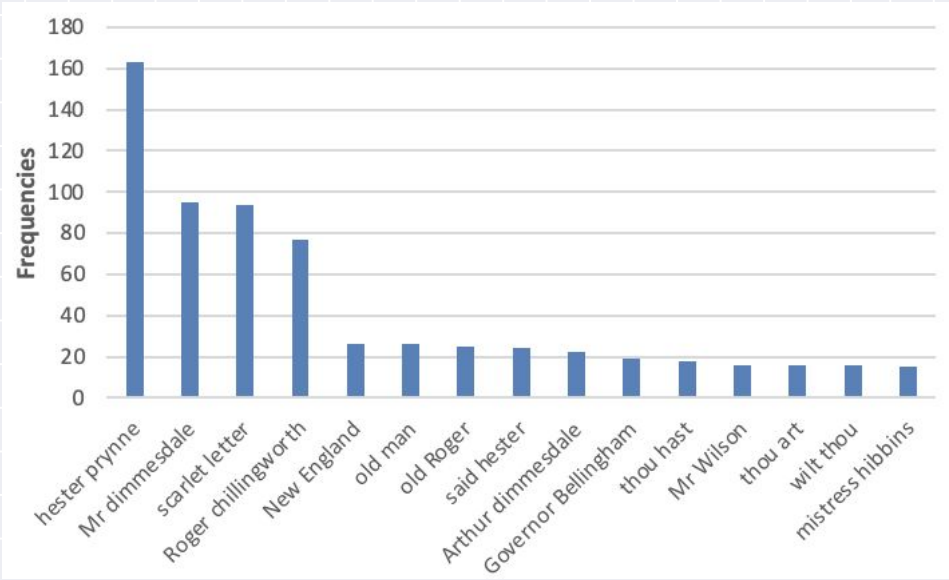
Similarities:

- Dialogue
- Character-driven narratives

Differences:

- Characters' interactions, context -specifics, and archaic language





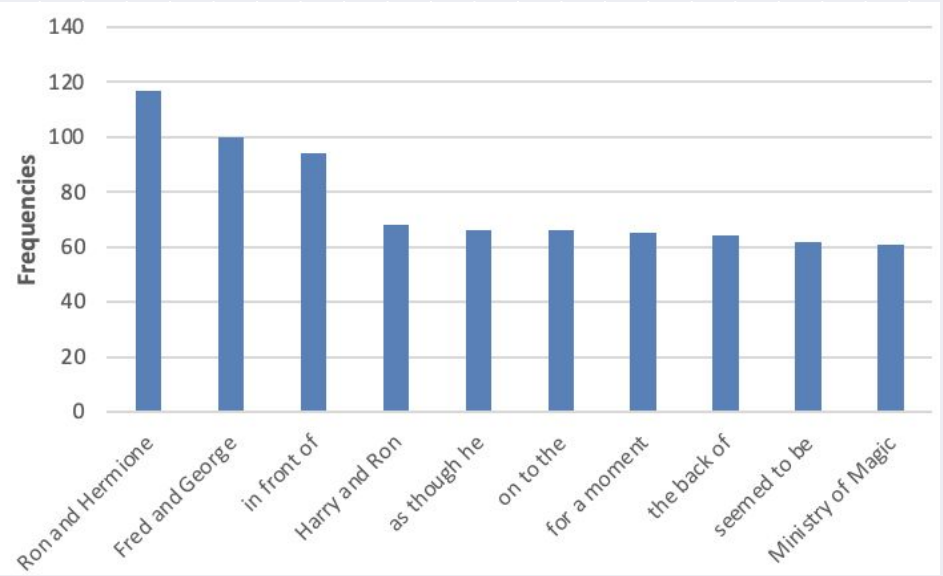
Harry Potter Book 5

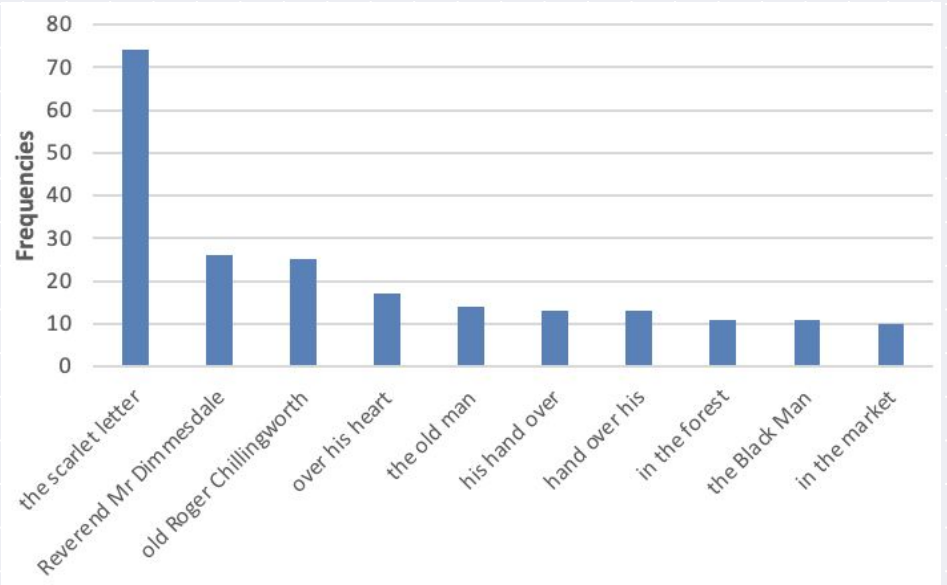
- "said Harry", "said Hermione", "said Ron"
- Characters: "Mrs. Weasley" and "Professor Umbridge"
- "Harry and" and "Ron and"

The Scarlet Letter

- "said Hester"
- Characters: "Hester Prynne", "Mr. Dimmesdale", "Roger Chillingworth", and "Governor Bellingham"
- Context-specifics: "New England" and "scarlet letter"
- "Thou hast" and "wilt thou"

N-grams — Trigrams





Harry Potter Book 5

- Trio's interactions: "Ron and Hermione" and "Harry and Ron"
- Other character relationships: "Fred and George"
- Entity: "Ministry of Magic"

The Scarlet Letter

- "Reverend Mr. Dimmesdale", "old Roger Chillingworth", and "the old man"
- Themes such as sin and public shaming
  - "The scarlet letter", "his hand over", "hand over his", "in the forest", and "in the market"

N-grams — Four-, Five-, and Six-grams

Harry Potter Book 5

- Organizations
  - “The Ministry of Magic”, “the Order of the Phoenix”, and “The Department of Mysteries”
- Class: “Defense Against the Dark Arts”
- Trio relationships: “Harry, Ron, and Hermione”

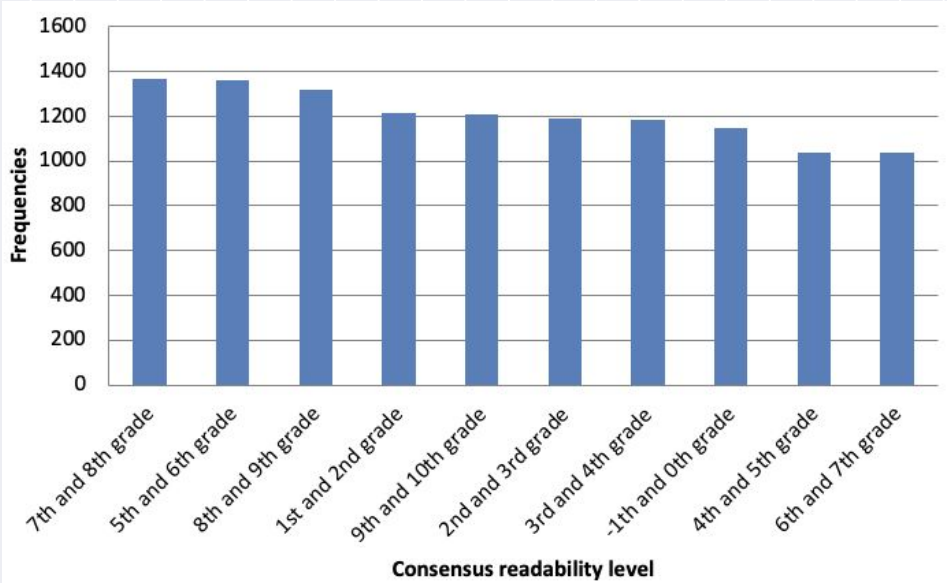
The Scarlet Letter

- The theme of sin and guilt: “hand over his heart” and “letter on her breast”

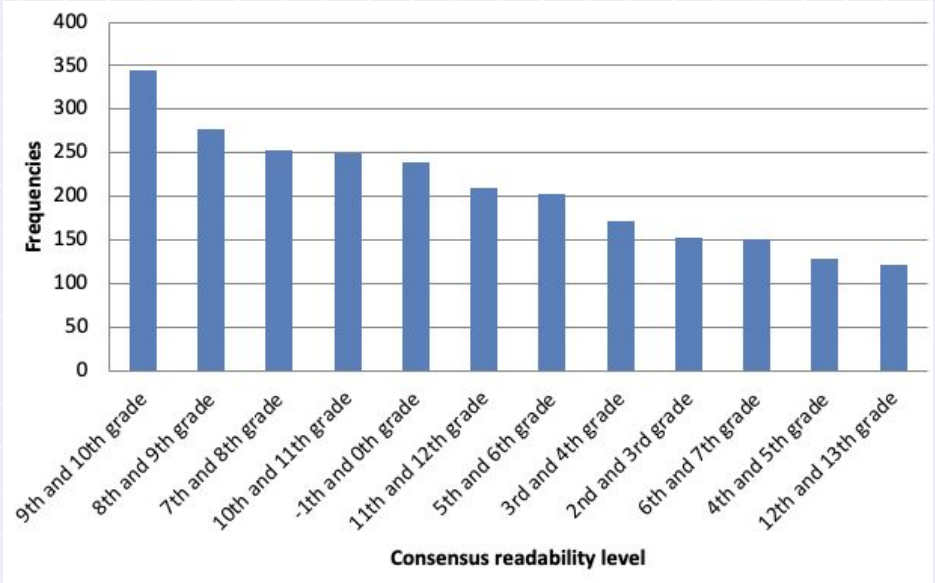
Text Readability

- Measure how easy a text is to read, based on complexity, familiarity, legibility, typography
- Consider factors such as sentence length, syllable density, word familiarity [2]
- NLP Suite computes the consensus readability level for each sentence with various readability metrics (Flesch-Kincaid Grade Level, SMOG, Fog Scale, etc.)

Text Readability



Harry Potter Book 5

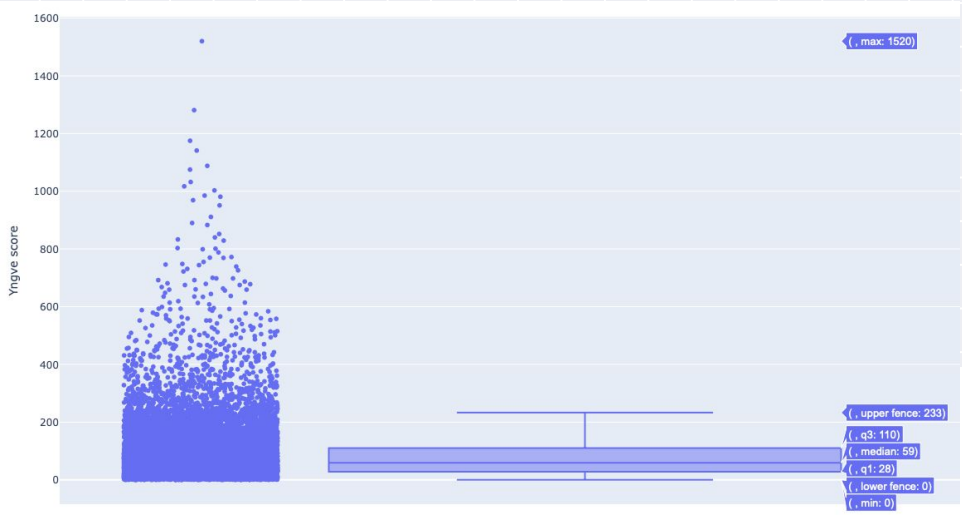


The Scarlet Letter

Sentence Complexity

- Measured by the number of subordinate clauses or clauses within a sentence
- Linked to how easily sentences are understood (text readability) and how accurately they are recalled [2]
- Yngve score is one of the ways to measure sentence complexity
  - The higher the score the more complex the sentence is

Sentence Complexity — Yngve score

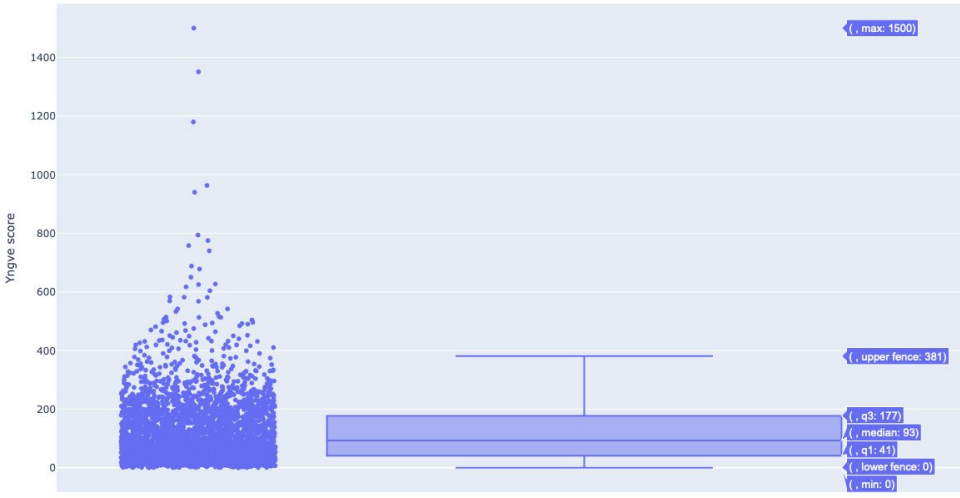


Harry Potter Book 5

- Max: 1520
- 3rd quartile: 110
- **Median: 59**
- 1st quartile: 28
- Min: 0

The Scarlet Letter

- Max: 1500
- 3rd quartile: 177
- **Median: 93**
- 1st quartile: 41
- Min: 0



Vocabulary Richness — Yule's K score

- Measure the text's lexical richness
  - The ratio of unique words to the total number of words
    - Higher ratio → more variety
  - How often words appear in the text
    - Many less common or specialized words → higher lexical richness
  - The use of longer or more complex words
- The larger the Yule's K value, the less diverse the vocabulary, and thus the easier the text

Vocabulary Richness — Yule's K score

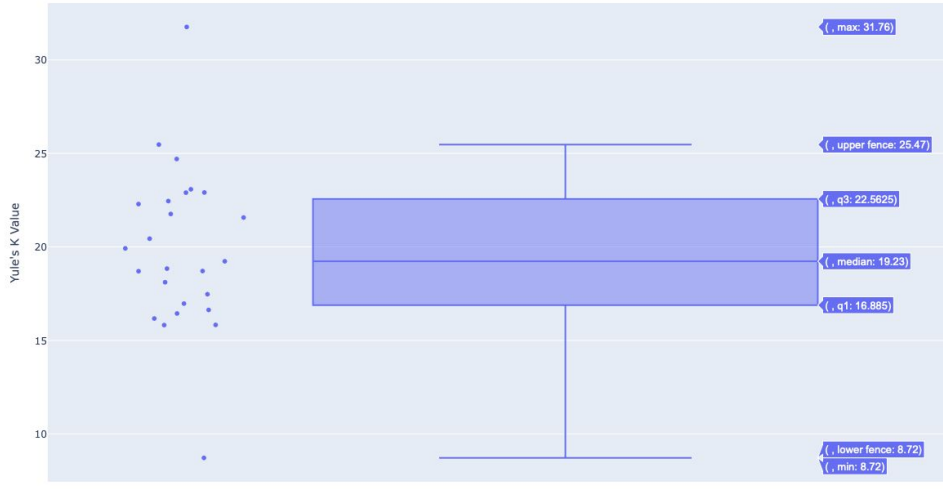
Harry Potter Book 5

- Max: 13.49
- 3rd quartile: 9.94
- **Median: 9.45**
- 1st quartile: 7.93
- Min: 4.63

Could due to its fantasy genre, which often introduces a broader array of topics and neologisms (such as magic spells)

The Scarlet Letter

- Max: 31.76
- 3rd quartile: 22.56
- **Median: 19.23**
- 1st quartile: 16.89
- Min: 8.72



Nominalization & Abstract/Concrete Words

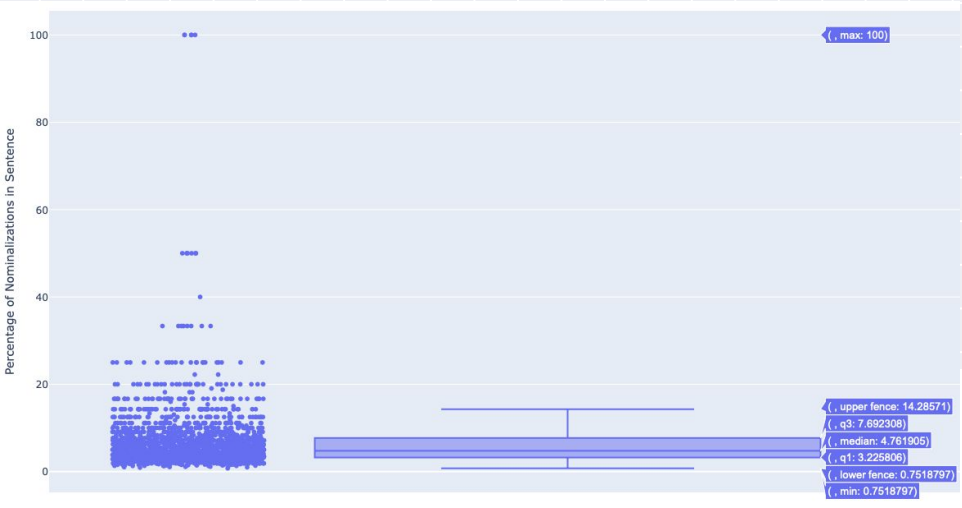
Nominalization is the process of changing a verb or adjective into a noun [3].

- E.g., “argue” → “argument”
- It make writing sound more formal
- It shift focus from the action (the verb) to the result or concept (the noun) and leads to more abstract sentences

Abstract/Concrete Words

- In contrast to abstract words, concrete words are psycholinguistically easier for people to remember [2].
- Concreteness is measure on a 5-point rating scale from 1 (abstract) to 5 (concreteness)
    - Concreteness measures the degree of a concept denoted by a word that refers to a perceptible entity.

Nominalization — Percentage of Nominalization in Sentence



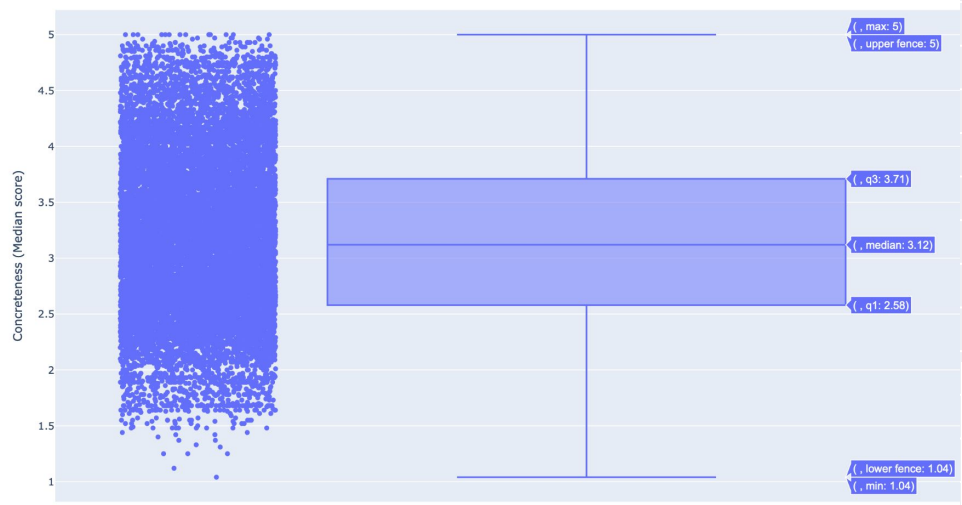
Harry Potter Book 5

- Max: 100
- 3rd quartile: 7.69
- **Median: 4.76**
- 1st quartile: 3.22
- Min: 0.75

The Scarlet Letter

- Max: 25
- 3rd quartile: 5
- **Median: 3.33**
- 1st quartile: 2.44
- Min: 0.96

Abstract/Concrete Vocabulary - Median Concreteness Values

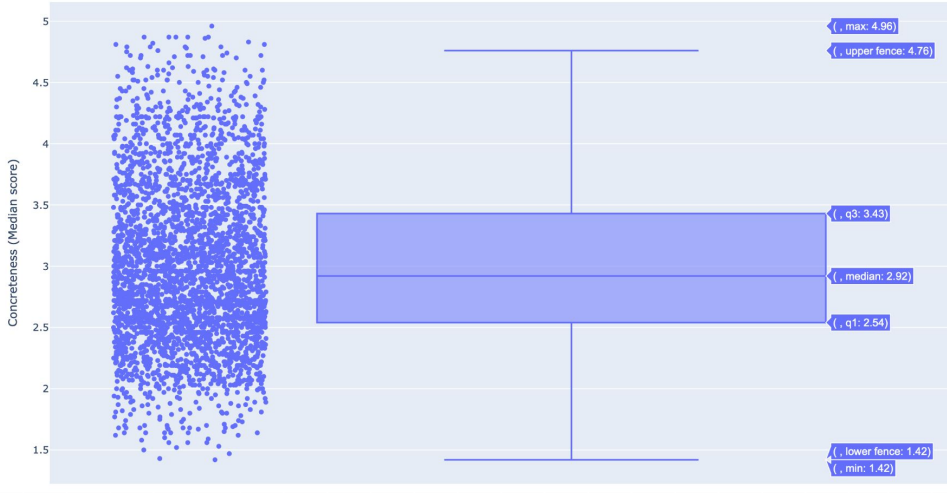


Harry Potter Book 5

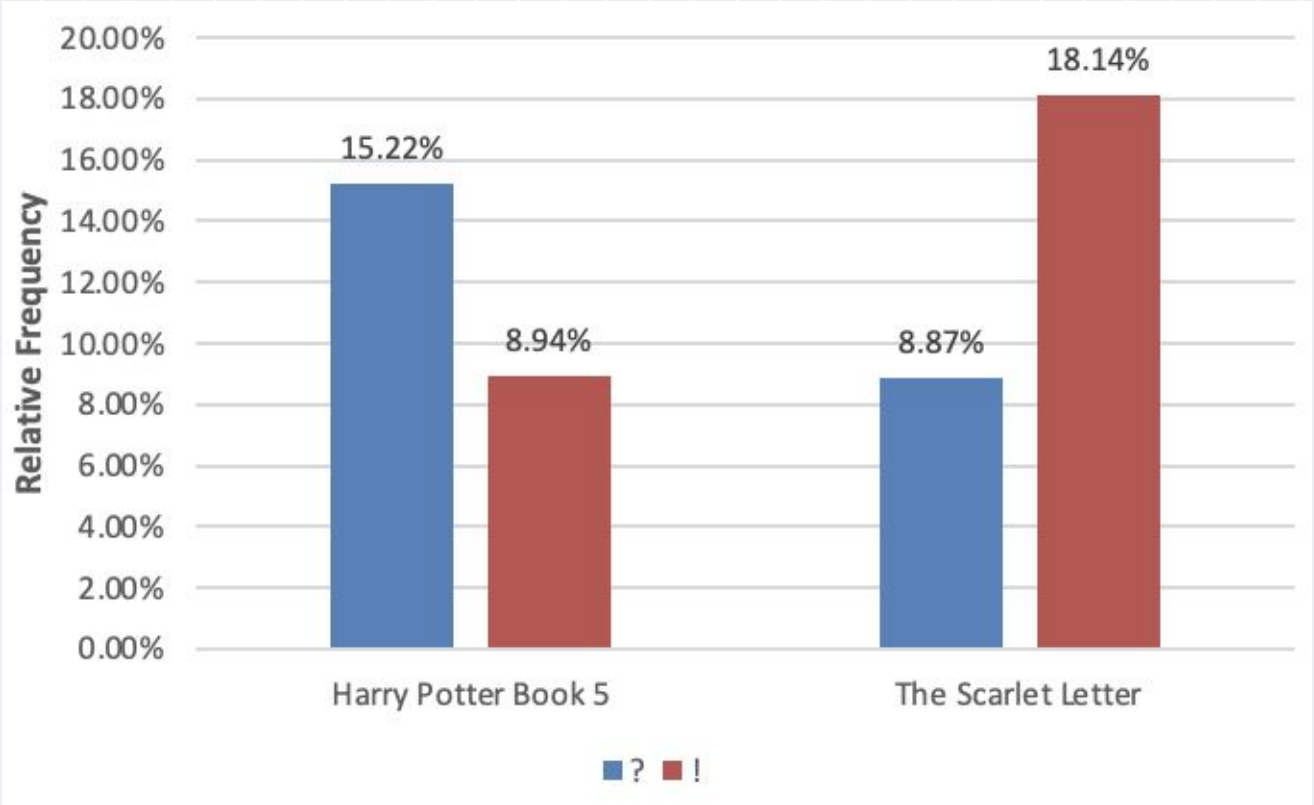
- Max: 5
- 3rd quartile: 3.71
- **Median: 3.12**
- 1st quartile: 2.58
- Min: 1.04

The Scarlet Letter

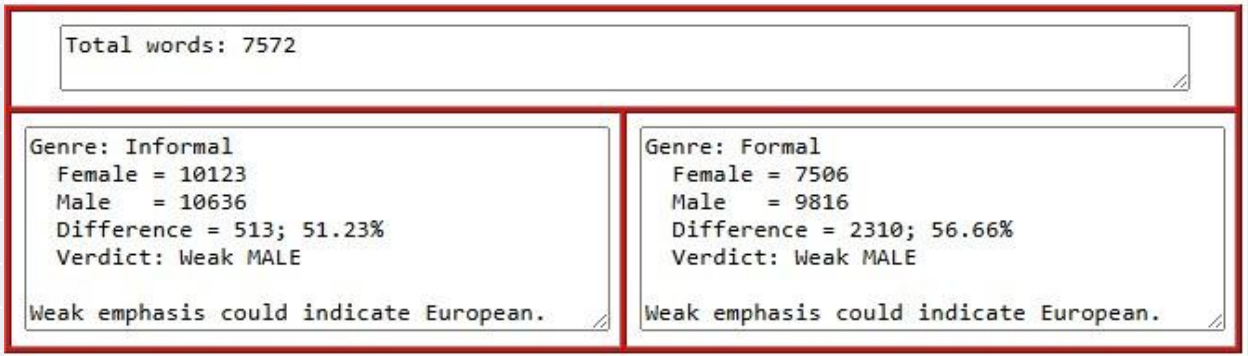
- Max: 4.96
- 3rd quartile: 3.43
- **Median: 2.92**
- 1st quartile: 2.54
- Min: 1.42



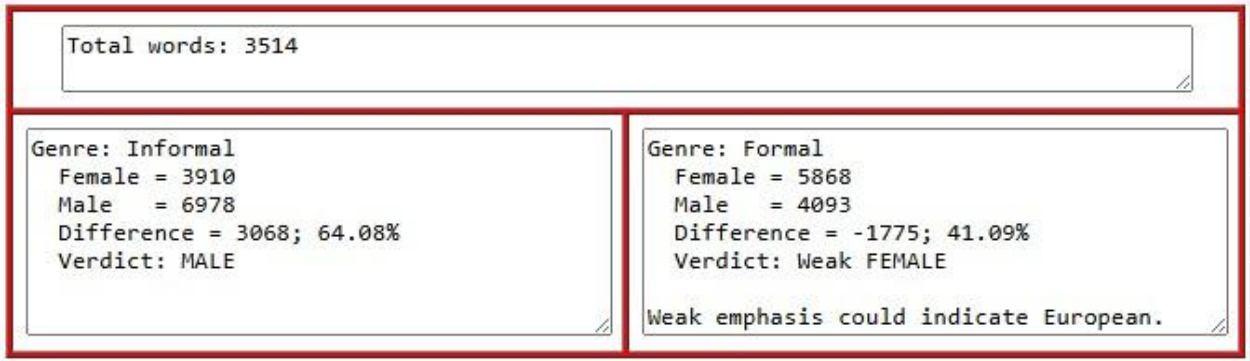
Punctuation as Figures of Pathos (? !)



GenderGuesser



Harry Potter Book 5



The Scarlet Letter

Topic Modeling

A technique that analyzes patterns in word usage across large sets of unstructured text data and identifies groups of words that frequently occur together, which represent a “topic”.

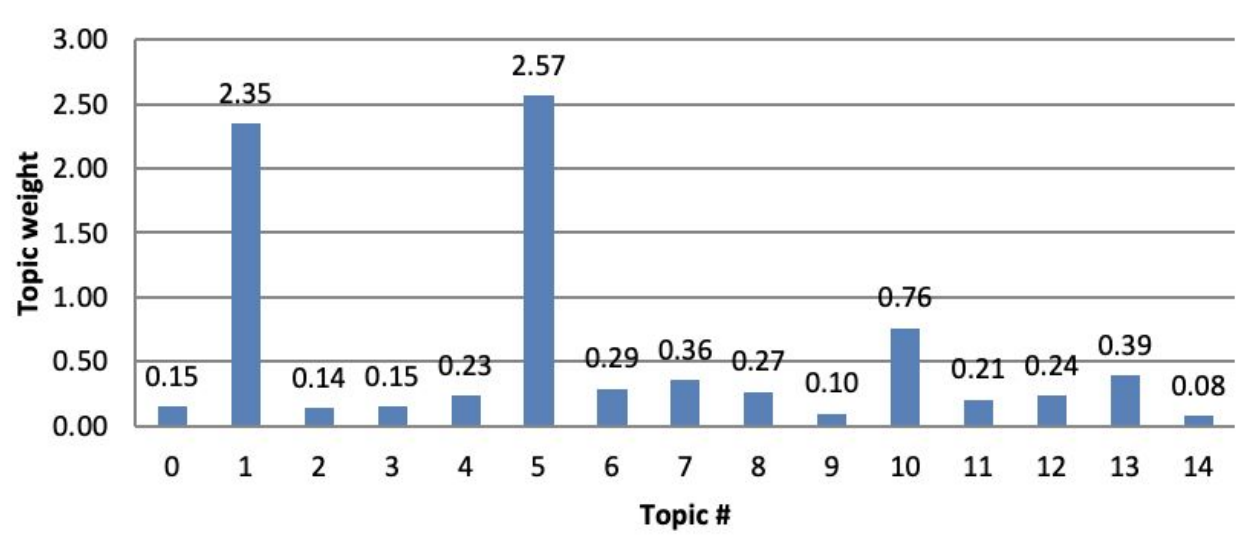
- These topics are not predefined — the computer does not know the meaning of each word — and are derived through statistical analysis [4].

Topic Modeling

MALLET	Gensim
• Fast and accurate but no visualization • To decide the number of topics, we should not end up with the majority of our original texts all in a very limited number of topics	• Not as accurate but nice visualization • To decide the number of topics, we want non-overlapping and large, spread-out circles on the intertopic distance map • The relevance metric (λ) helps determine how important different words are for each topic • Small λ (near 0) → rare, but exclusive terms for the selected topic • Large λ (near 1) → frequent, but not necessarily exclusive, terms for the selected topic.

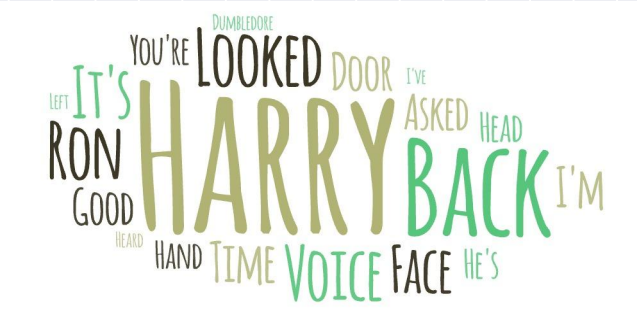
MALLET Topic Modeling — Harry Potter Book 5

We decided to use 15 topics for MALLET topic modeling because it has two high-weight topics and only one low-weight topic (below 0.1).



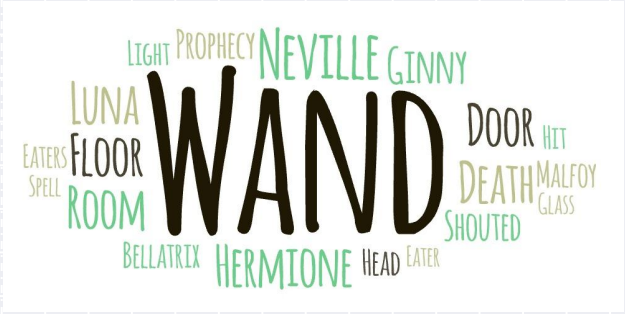
MALLET Topic Modeling — Harry Potter Book 5

Topic 1 (weight = 2.35)



Topic 5 (weight = 2.57)

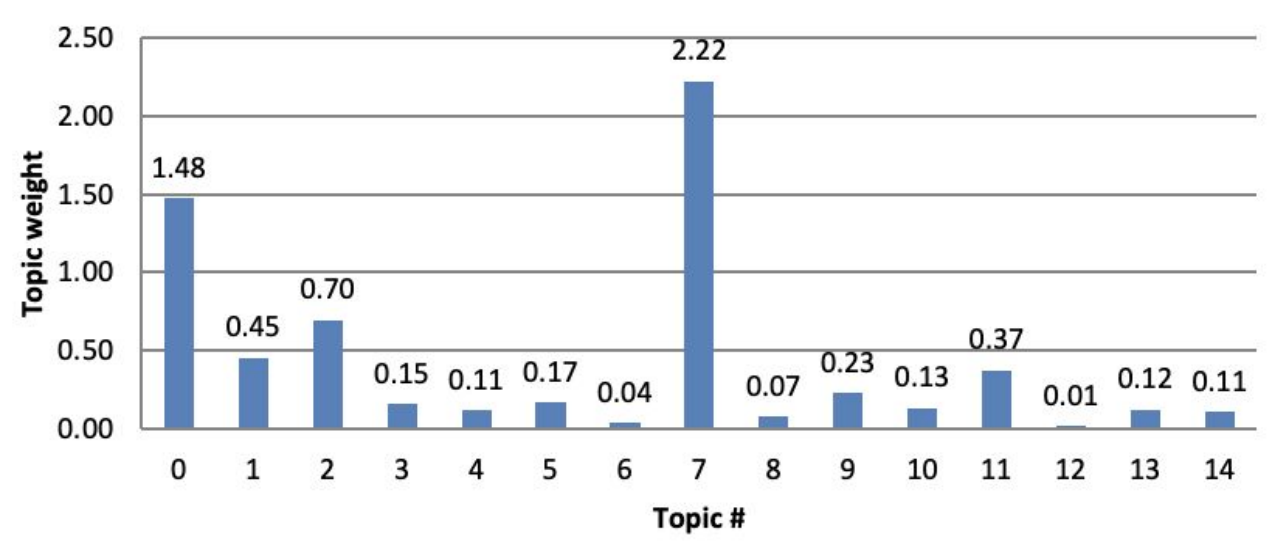
Topic 10 (weight = 0.76)



Topic 13 (weight = 0.39)

MALLET Topic Modeling — The Scarlet Letter

We decided to use 15 topics for MALLET topic modeling because it has two high-weight topics and three low-weight topics (below 0.1).

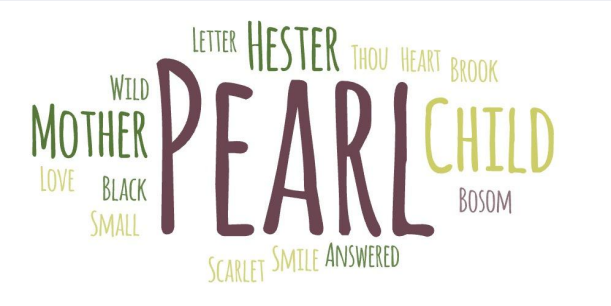


MALLET Topic Modeling — The Scarlet Lett

Topic 0 (weight = 1.48)

Topic 7 (weight = 2.22)

Topic 9 (weight = 0.23)



Topic 2 (weight = 0.69)

Gensim Topic Modeling — Harry Potter Book 5

We decided to use 8 topics with lambda = 0.5.

file:///Users/zijingzhou/Desktop/Emory/QTM_446W/Output/TM-Gensim_Harry_Potter_Book_5-all/TM-Gensim_Harry_Potter_Book_5-8/NLP_Gensim_topic_modeling_Dir_Harry_Potter_Book_5.html#topic=0&lambda=1&term=

Gensim Topic Modeling — The Scarlet Letter

We decided to use 10 topics with lambda = 0.8.

file:///Users/zijingzhou/Desktop/Emory/QTM_446W/Output/TM-Gensim_The_Scarlet_Letter-10/NLP_Gensim_topic_modeling_Dir_The_Scarlet_Letter.html

Word2Vector

- Word2Vec is a machine learning algorithm used to generate word embeddings.
- Word embeddings are dense vector representations of words in a continuous vector space.
- Word2Vec takes a text corpus as input and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space.
- Word2Vec in the NLP Suite can be done using BERT or Gensim, with Gensim having two different learning architectures that it can use.
- Word2Vec can be used to perform a variety of NLP tasks, such as finding synonyms, calculating analogies, and classifying documents by topic.

Gensim vs BERT

Gensim Word2Vec:

- Shallow, two-layer neural network.
- Context-independent: Each word has a single vector representation, regardless of its context.
- Relatively fast and efficient to train.
- Good for capturing general semantic relationships between words.

BERT Word Embeddings:

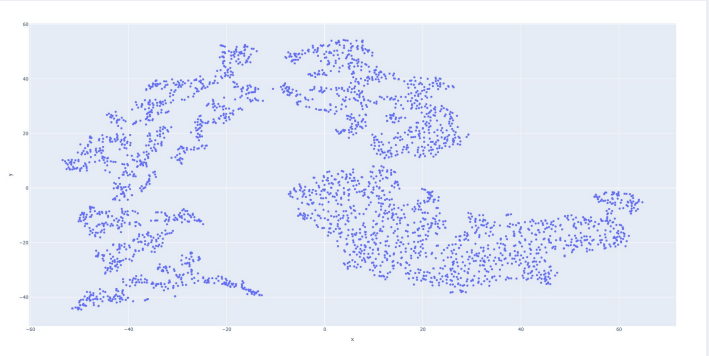
- Deep, multi-layered transformer network.
- Context-dependent: Each word can have different vector representations depending on its surrounding words.
- More computationally expensive to train.
- Better at capturing nuanced meanings and relationships between words in specific contexts.

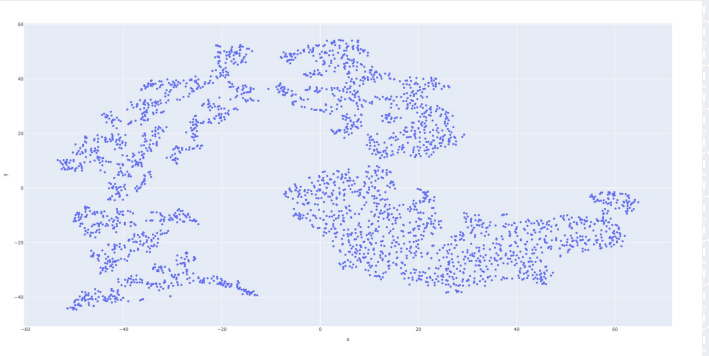
Gensim Word Vector Plots

CBOW

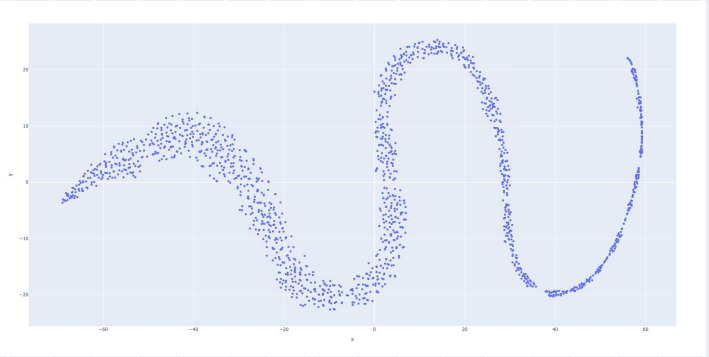
Skip-Gram

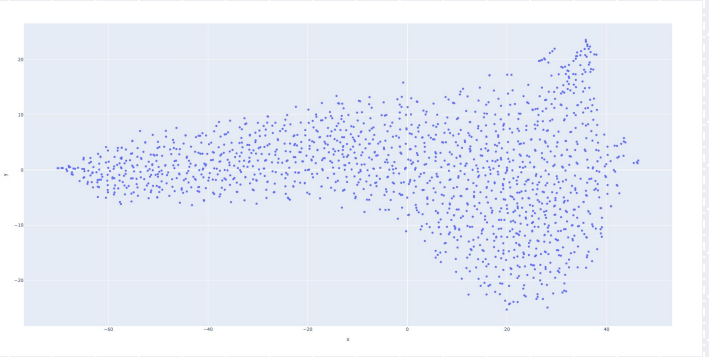
Harry Potter 5





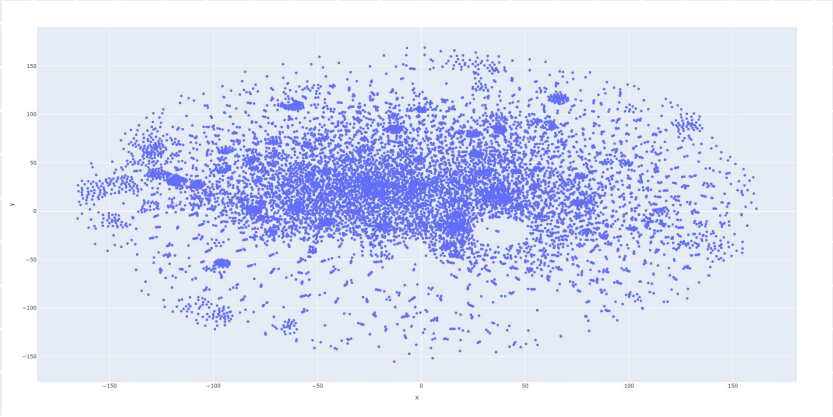
Scarlet Letter



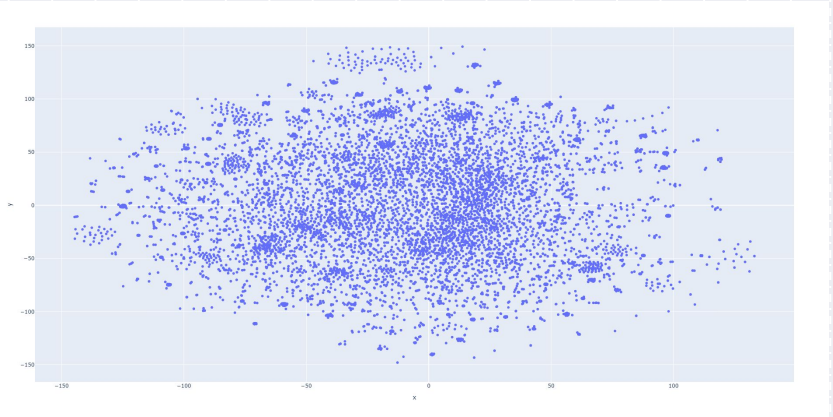


BERT Word Vector Plots

Harry Potter 5



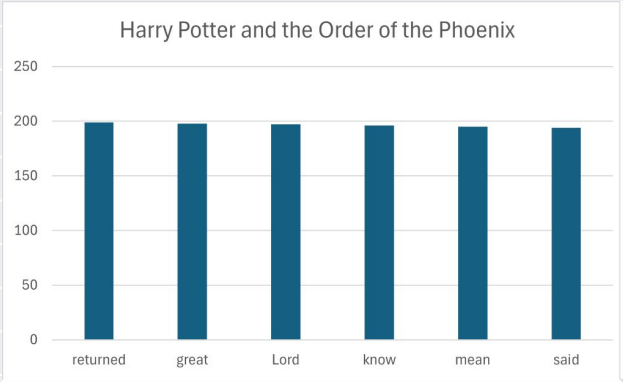
Scarlet Letter

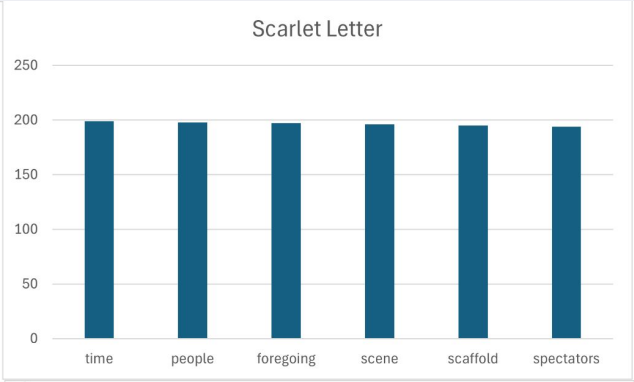


Finding Central Words

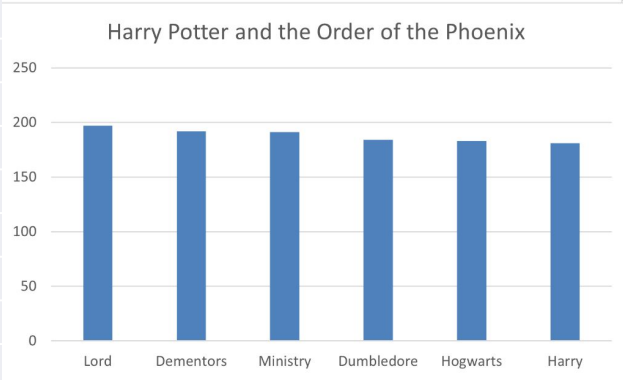
Frequency of appearance in the Word_1 column of the BERT output was used to determine words with high centrality in word associations

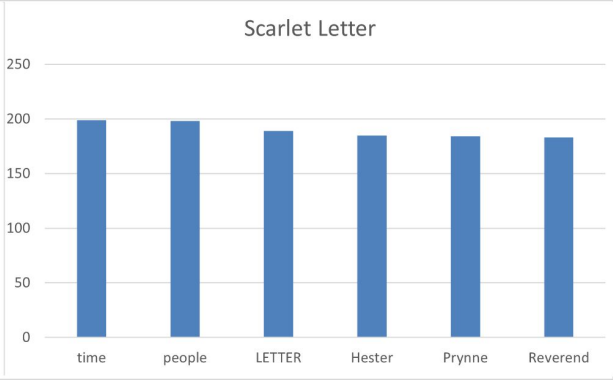
Highest Frequency



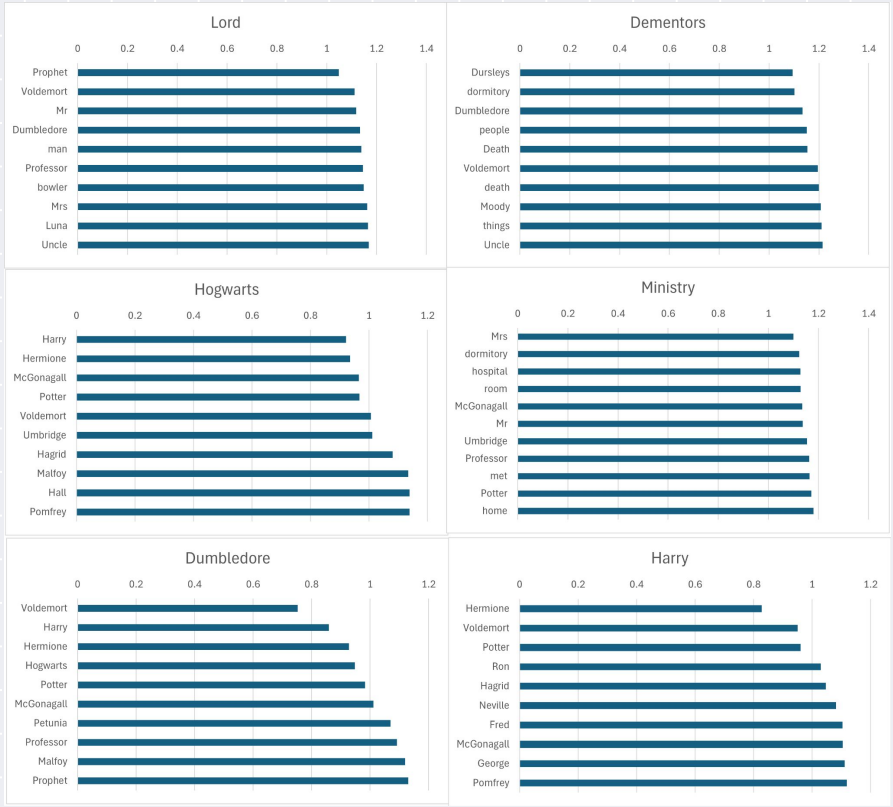


Custom Choice of High Frequency Nouns





Closest Words using Euclidean Distance in Harry Potter 5

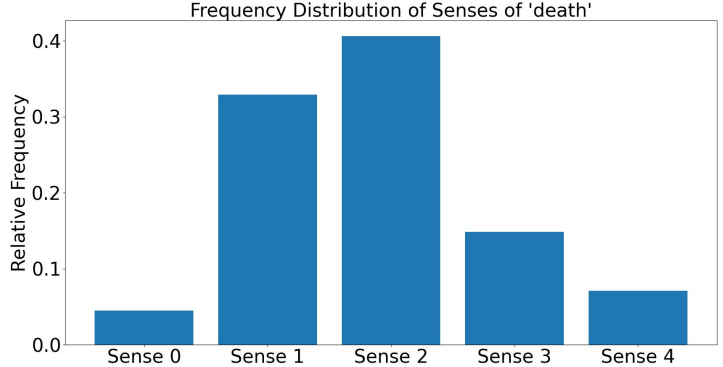


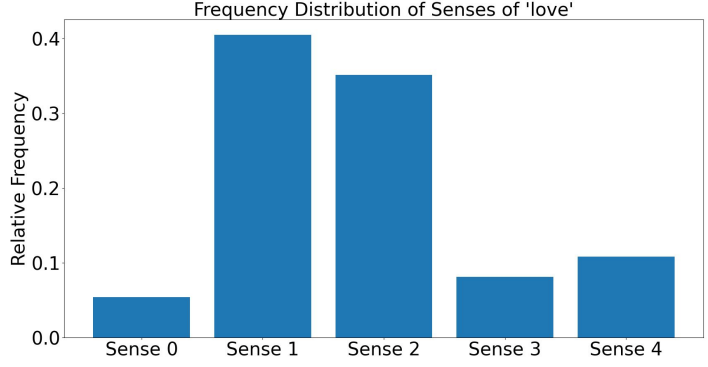
Closest Words using Euclidean Distance in Scarlet Letter

Word Sense Induction (WSI)

Word Sense Induction (WSI):

- Automatically identifies different meanings (senses) of a word based on its context.
- Helps disambiguate polysemous words (words with multiple meanings).
- Uses word embeddings and clustering algorithms to group different uses of a word.
- Provides insights into the nuanced use of language in a corpus.

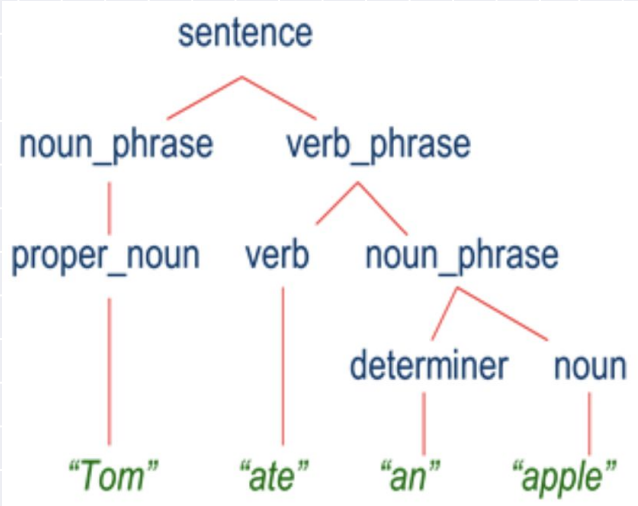




Parsers

A parser can determine the syntactic structure of a text by analyzing its constituent words based on an underlying grammar of the language of the text (e.g., English).

- For instance, it would figure out the structure of the sentence “Tom ate an apple” assigning each word in the sentence its proper syntactic label [2].



Compare between parsers

Stanford CoreNLP
✅ Strong performance on formal text, particularly well-formed sentences.
✅ High accuracy in dependency parsing.
❌ Slower processing speed.
❌ High memory usage.

spaCy
✅ Fastest processing time (ideal for large corpora).
✅ Efficient memory usage.
❌ Lower accuracy on complex sentence structures.
❌ Less robust to noisy/unstructured text.

Stanza
✅ Best accuracy due to neural-based approach.
✅ Handles noisy/unstructured data well.
❌ Slowest processing speed.
❌ Higher computational demands.

Stanford CoreNLP

9 minutes to run
High memory

ID	Form	Lemma	POS	NER	Head	DepRel	Deps	Clause Tag	Record ID	Sentence ID	Document ID	Document
1	It	it	PRP	O	5	nsubj	5:nsubj		1	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
2	is	be	VBZ	O	5	cop	5:cop		2	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
3	a	a	DT	O	4	det	4:det		3	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
4	little	little	JJ	O	5	obl:npmod	5:obl:npmod		4	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
5	remarkable	remarkable	JJ	O	0	ROOT	0:ROOT		5	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
6	,	,	,	O	5	punct	5:punct		6	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
7	that	that	IN	O	10	mark	10:mark		7	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
8	,Äî	,Äî	HYPH	O	10	punct	10:punct		8	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
9	though	though	RB	O	10	advmod	10:advmod		9	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
10	disinclined	disinclined	JJ	O	5	ccomp	5:ccomp		10	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
11	to	to	TO	O	12	mark	12:mark		11	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
12	talk	talk	VB	O	10	xcomp	10:xcomp		12	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
13	overmuch	overmuch	IN	O	15	case	15:case		13	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
14	of	of	IN	O	15	case	15:case		14	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
15	myself	myself	PRP	O	12	obl:of	12:obl:of		15	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
16	and	and	CC	O	18	cc	18:cc		16	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
17	my	my	PRP$	O	18	nmod:poss	18:nmod:poss		17	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
18	affairs	affair	NNS	O	15	conj:and	12:obl:of|15:conj:and		18	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
19	at	at	IN	O	21	case	21:case		19	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
20	the	the	DT	O	21	det	21:det		20	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
21	fireside	fireside	NN	O	12	obl:at	12:obl:at		21	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
22	,	,	,	O	12	punct	12:punct		22	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
23	and	and	CC	O	27	cc	27:cc		23	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
24	to	to	IN	O	27	case	27:case		24	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
25	my	my	PRP$	O	27	nmod:poss	27:nmod:poss		25	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
26	personal	personal	JJ	O	27	amod	27:amod		26	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
27	friends	friend	NNS	O	12	conj:and	10:xcomp|12:conj:and		27	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
28	,Äî	,Äî	:	O	27	punct	27:punct		28	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
29	an	a	DT	O	31	det	31:det		29	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
30	autobiograph	autobiograph	JJ	O	31	amod	31:amod		30	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
31	impulse	impulse	NN	O	38	nsubj	38:nsubj		31	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
32	should	should	MD	O	38	aux	38:aux		32	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
33	twice	twice	RB	O	38	advmod	38:advmod		33	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt

spaCy

5 minutes to run

ID	Form	Lemma	POS	NER	Multi-Word E	Head	DepRel	Sentence ID	Sentence	Document ID	Document
0	Ôªølt	Ôªølt	PUNCT		O	1	punct	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
1	is	be	AUX		O	1	ROOT	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
2	a	a	DET		O	3	det	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
3	little	little	ADJ		O	4	npadvmod	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
4	remarkable	remarkable	ADJ		O	1	acomp	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
5	,	,	PUNCT		O	9	punct	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
6	that	that	ADV		O	9	advmod	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
7	‚Äî	‚Äî	PUNCT		O	9	punct	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
8	though	though	SCONJ		O	9	mark	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
9	disinclined	disinclined	ADJ		O	1	advcl	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
10	to	to	PART		O	11	aux	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
11	talk	talk	VERB		O	9	xcomp	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
12	overmuch	overmuch	ADJ		O	11	prep	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
13	of	of	ADP		O	12	prep	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
14	myself	myself	PRON		O	13	pobj	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
15	and	and	CCONJ		O	9	cc	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
16	my	my	PRON		O	17	poss	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
17	affairs	affair	NOUN		O	9	conj	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
18	at	at	ADP		O	17	prep	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
19	the	the	DET		O	20	det	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
20	fireside	fireside	NOUN		O	18	pobj	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
21	,	,	PUNCT		O	9	punct	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
22	and	and	CCONJ		O	9	cc	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
23	to	to	ADP		O	37	prep	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
24	my	my	PRON		O	26	poss	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
25	personal	personal	ADJ		O	26	amod	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
26	friends	friend	NOUN		O	23	pobj	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
27	‚Äî	‚Äî	PUNCT		O	37	punct	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
28	an	an	DET		O	30	det	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
29	autobiograph	autobiograph	ADJ		O	30	amod	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
30	impulse	impulse	NOUN		O	37	nsubj	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
31	should	should	AUX		O	37	aux	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
32	twice	twice	ADV		O	37	advmod	1	Ôªølt is a littl	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt

Stanza

5 minutes to run

ID	Form	Lemma	POS	NER	feats	Multi-Word E	Head	DepRel	Record ID	Sentence ID	Document ID	Document
1	Ôªø	Ôªø	PUNCT	O			6	punct	1	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
2	It	it	PRON	O	Case=Nom|Gender=Neut		6	nsubj	2	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
3	is	be	AUX	O	Mood=Ind|Number=Sing|I		6	cop	3	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
4	a	a	DET	O	Definite=Ind|PronType=Ar		5	det	4	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
5	little	little	ADJ	O	Degree=Pos		6	obl:npmod	5	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
6	remarkable	remarkable	ADJ	O	Degree=Pos		0	root	6	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
7	,	,	PUNCT	O			8	punct	7	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
8	that	that	SCONJ	O			11	reparandum	8	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
9	,Äî	,Äî	PUNCT	O			8	punct	9	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
10	though	though	SCONJ	O			11	mark	10	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
11	disinclined	disinclined	ADJ	O	Degree=Pos		6	advcl	11	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
12	to	to	PART	O			13	mark	12	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
13	talk	talk	VERB	O	VerbForm=Inf		11	xcomp	13	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
14	overmuch	overmuch	ADV	O			13	advmod	14	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
15	of	of	ADP	O			16	case	15	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
16	myself	myself	PRON	O	Case=Acc|Number=Sing|		13	obl	16	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
17	and	and	CCONJ	O			19	cc	17	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
18	my	my	PRON	O	Case=Gen|Number=Sing|		19	nmod:poss	18	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
19	affairs	affair	NOUN	O	Number=Plur		16	conj	19	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
20	at	at	ADP	O			22	case	20	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
21	the	the	DET	O	Definite=Def|PronType=Ar		22	det	21	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
22	fireside	fireside	NOUN	O	Number=Sing		13	obl	22	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
23	,	,	PUNCT	O			24	punct	23	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
24	and	and	CCONJ	O			39	cc	24	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
25	to	to	ADP	O			28	case	25	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
26	my	my	PRON	O	Case=Gen|Number=Sing|		28	nmod:poss	26	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
27	personal	personal	ADJ	O	Degree=Pos		28	amod	27	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
28	friends	friend	NOUN	O	Number=Plur		39	obl	28	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
29	,Äî	,Äî	PUNCT	O			28	punct	29	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
30	an	a	DET	O	Definite=Ind|PronType=Ar		32	det	30	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
31	autobiograph	autobiograph	ADJ	O	Degree=Pos		32	amod	31	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
32	impulse	impulse	NOUN	O	Number=Sing		39	nsubj	32	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt
33	should	should	AUX	O	VerbForm=Fin		39	aux	33	1	1	/Users/zijingzhou/Desktop/Emory/QTM_446W/The_Scarlet_Letter/Hawthorne_The_Scarlet_Letter_01.txt

Measuring the accuracy

Labeled Attachment Score (LAS):

$$LAS = \frac{\text{Number of correctly predicted head + label pairs}}{\text{Total words in corpus}}$$

High LAS (~90%) → Good syntactic accuracy
Low LAS (~50-60%) → Parser struggles with certain structures

Summary

	Harry Potter Book 5	The Scarlet Letter
Sentence length	Shorter	Longer
Text readability/ sentence complexity	More readable and less complex	Less readable and more complex
Vocabulary richness	More diverse	Less diverse
Nominalization	Higher percent normalized words per sentence	Lower percent normalized words per sentence
Abstract/concrete words	More concrete words	Less concrete words
Punctuation (?!)	More use of ?	More use of !
Similar themes	Character-driven and dialogue-focused	
Corpus-specific theme	Magic, school, battles, different groups and organizations, etc.	Sin, guilt, repentance, motherhood, public shaming, etc.

Works Cited

[1] Kumar, Prachi. “An Introduction to N-Grams: What Are They and Why Do We Need Them?” XRDS, 21 Oct. 2017, https://blog.xrds.acm.org/2017/10/introduction-n-grams-need/.

[2] Franzosi, Roberto. NLP TIPS files.

[3] Franzosi, Roberto, et al. “Ways of Measuring Agency.” Sociological Methodology, vol. 42, no. 1, Aug. 2012, pp. 1–42, https://doi.org/10.1177/0081175012462370.

[4] Graham, Shawn, et al. “Getting Started with Topic Modeling and MALLET.” Programming Historian, 2 Sept. 2012, programminghistorian.org/en/lessons/topic-modeling-and-mallet.

Granger, J. (2008). Harry Potter and the Deathly Hallows: Unlocking the Secrets of the Final Book in the Harry Potter Series. Tyndale House Publishers.

Mendelsohn, D. (2018). The Odyssey and Harry Potter: A Study in Character Perspective. The New Yorker. Available at: https://www.newyorker.com

Pottermore (2018). How long is each Harry Potter book? Available at: https://www.pottermore.com

Publishers Weekly (2003). Review: Harry Potter and the Order of the Phoenix. Available at: https://www.publishersweekly.com

The Guardian (2016). How the Harry Potter books grew up with their readers. Available at: https://www.theguardian.com

Tiffin, J. & Dowling, C. (2010). The Narrative Voice in Harry Potter: Third-Person Limited and Beyond. Children’s Literature Review, 35(2), pp. 145-162.

American Library Association (2019). Frequently Taught American Literature in Schools. Available at: https://www.ala.org

Bercovitch, S. (1991). The Office of The Scarlet Letter. Baltimore: Johns Hopkins University Press.

College Board (2021). AP English Literature and Composition: Required Reading List. Available at: https://www.collegeboard.org

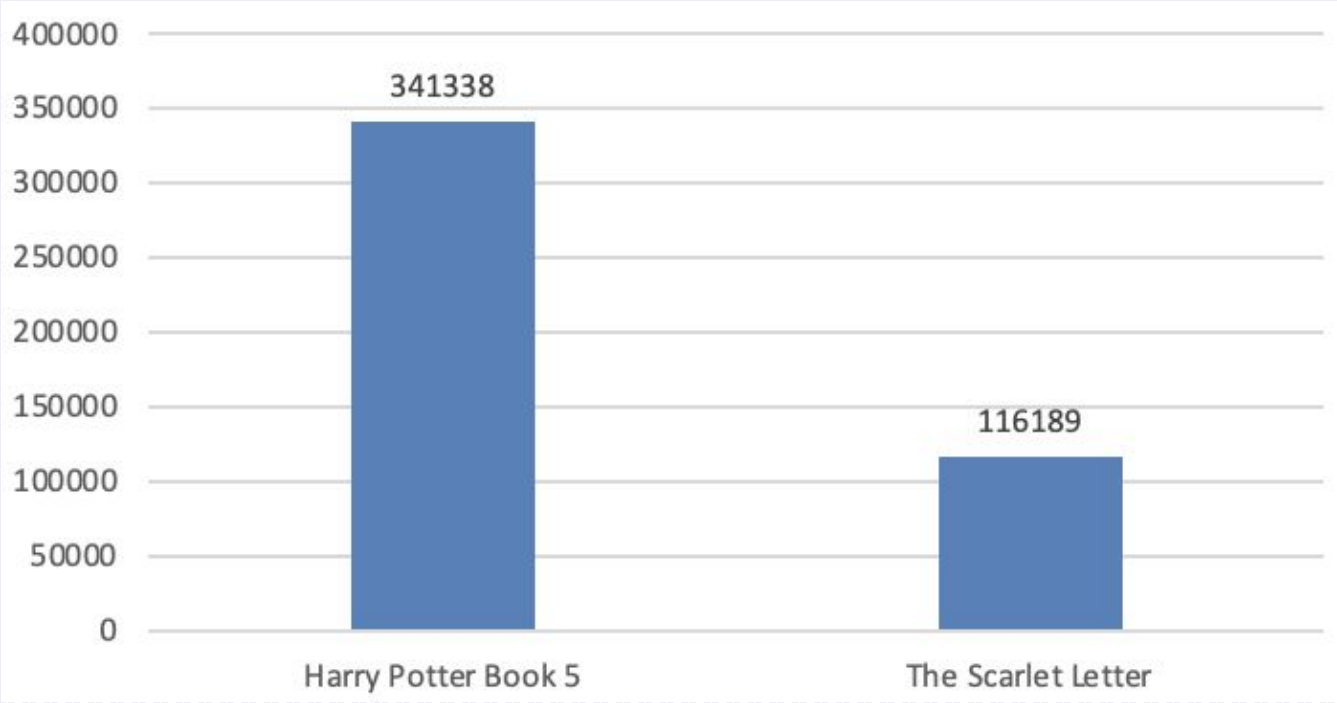
Hawthorne, N. (1850). The Scarlet Letter. Ticknor, Reed, and Fields.

Person, L.S. (2007). The Cambridge Introduction to Nathaniel Hawthorne. Cambridge University Press.

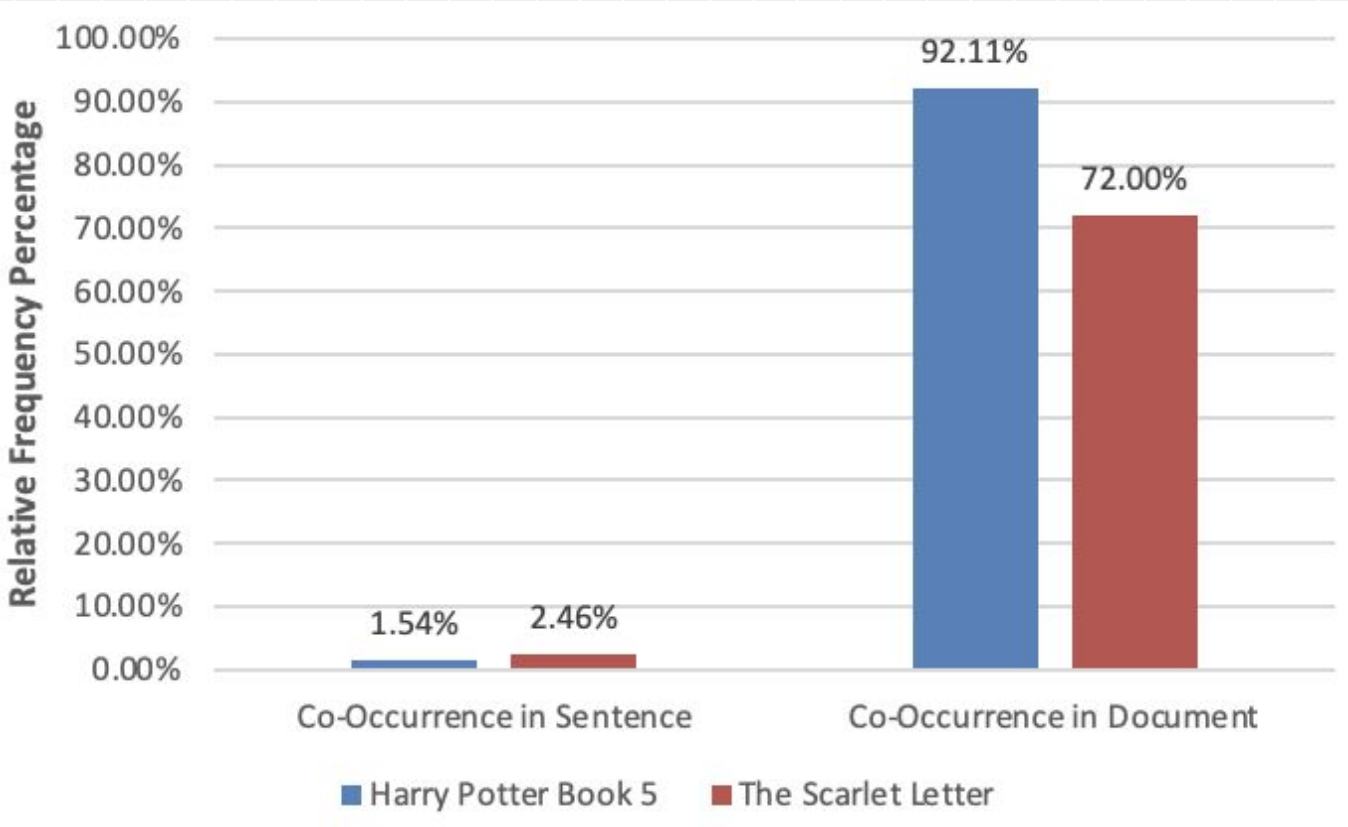
Reynolds, D.S. (2008). Beneath the American Renaissance: The Subversive Imagination in the Age of Emerson and Melville. Harvard University Press.

Extra Slides

Corpus Statistics — Total Syllable Count

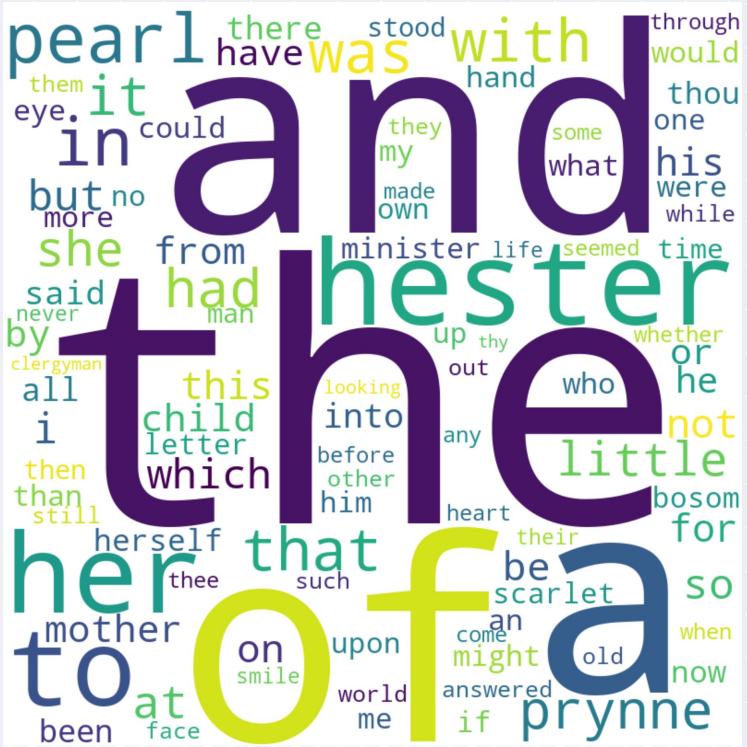


Co-occurrences VIEWER



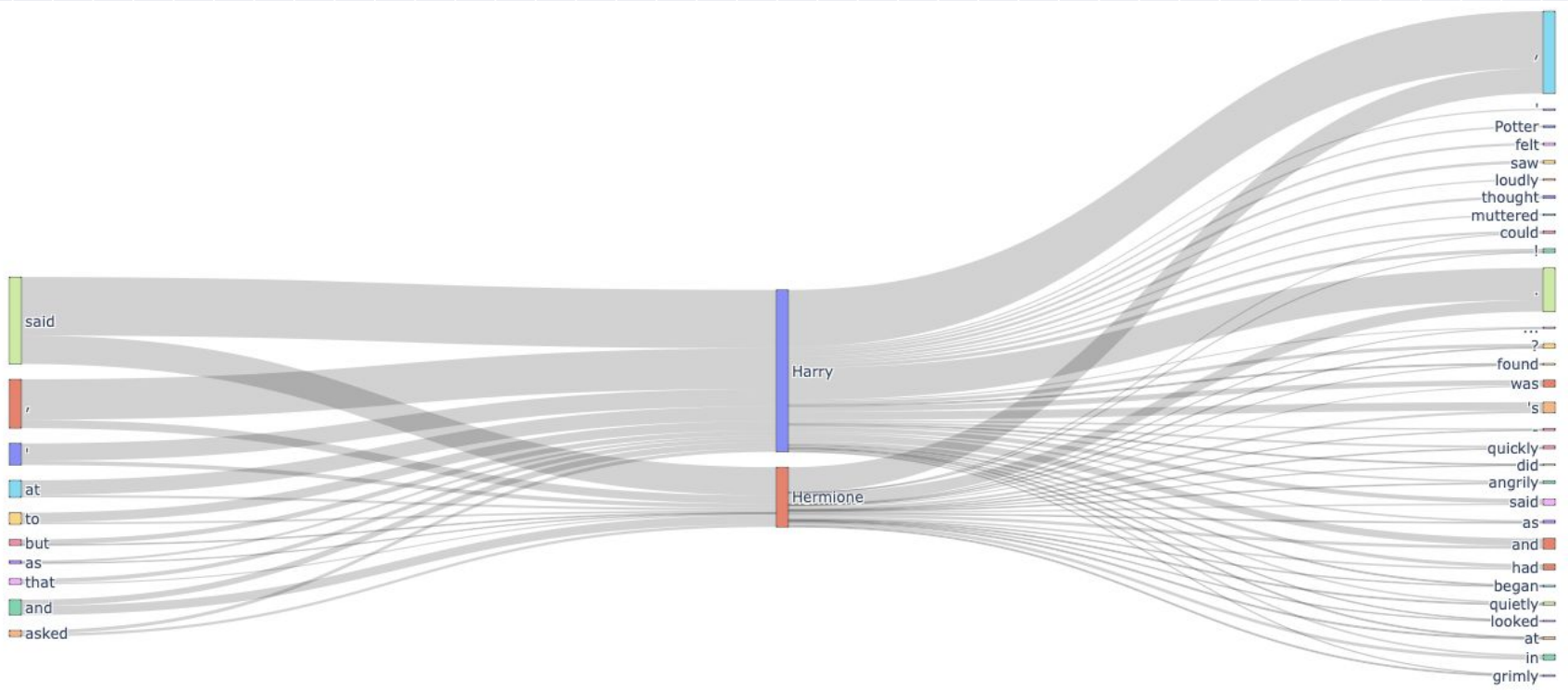
Words/Collocations Searches - Word

Harry Potter Book 5: “Harry” and “Hermione”



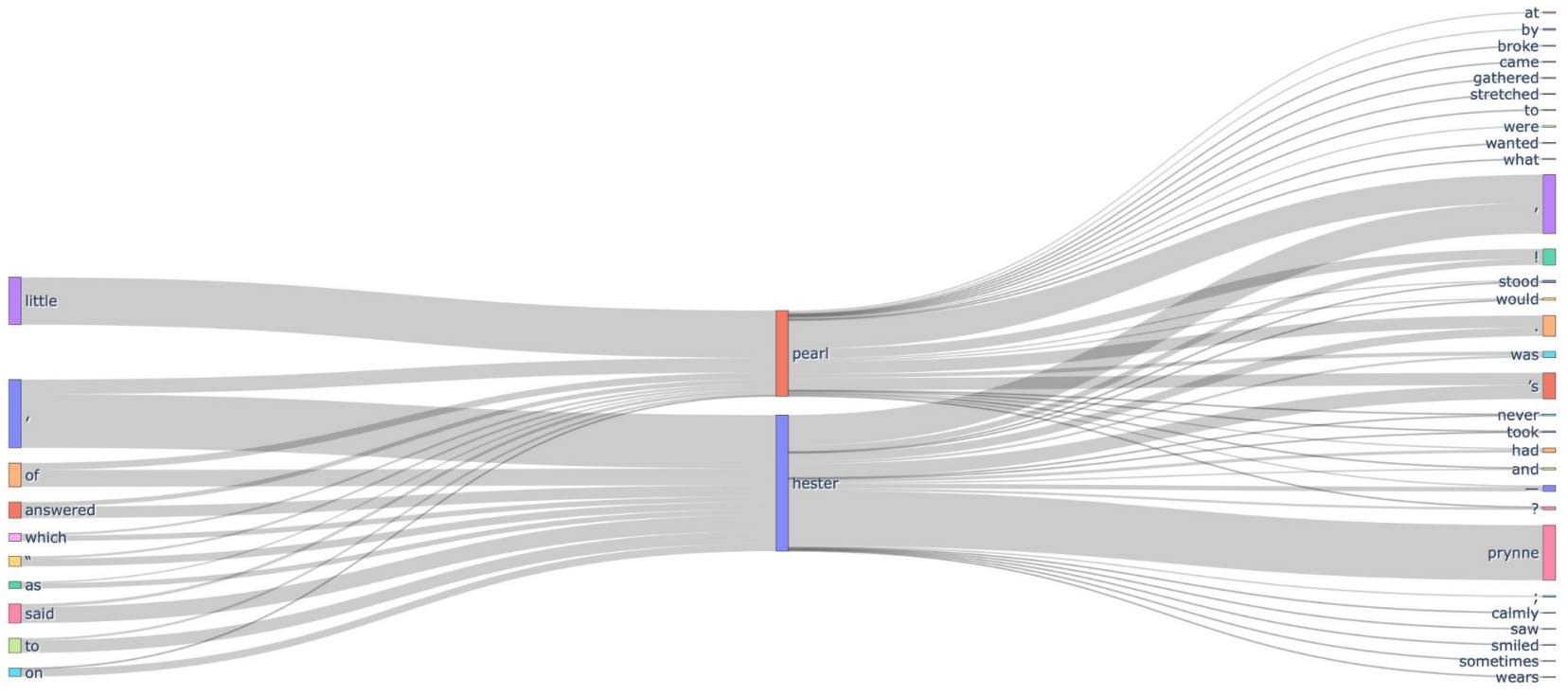
The Scarlet Letter: “Hester” and “Pearl”

Words/Collocations Searches - Sankey



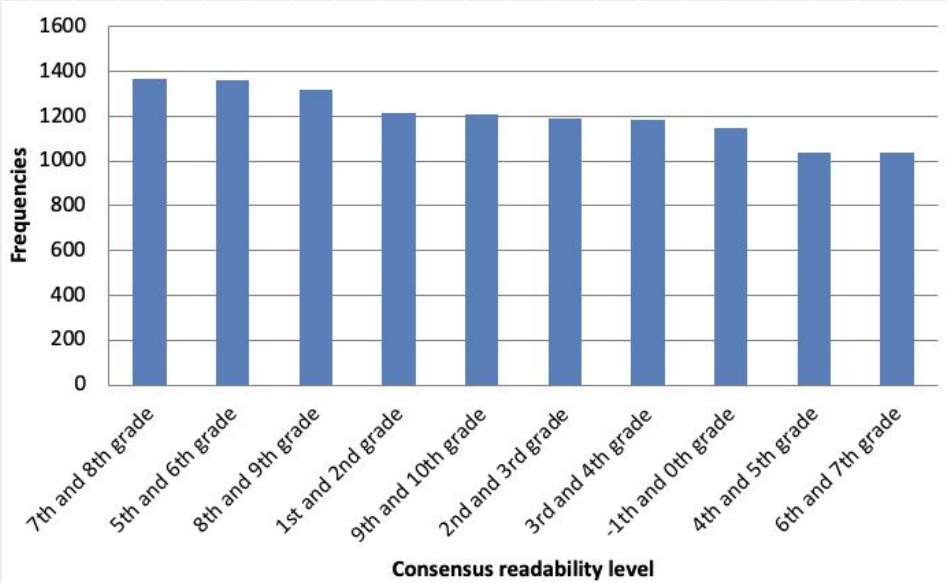
Harry Potter Book 5: “Harry” and “Hermione”

Words/Collocations Searches - Sankey

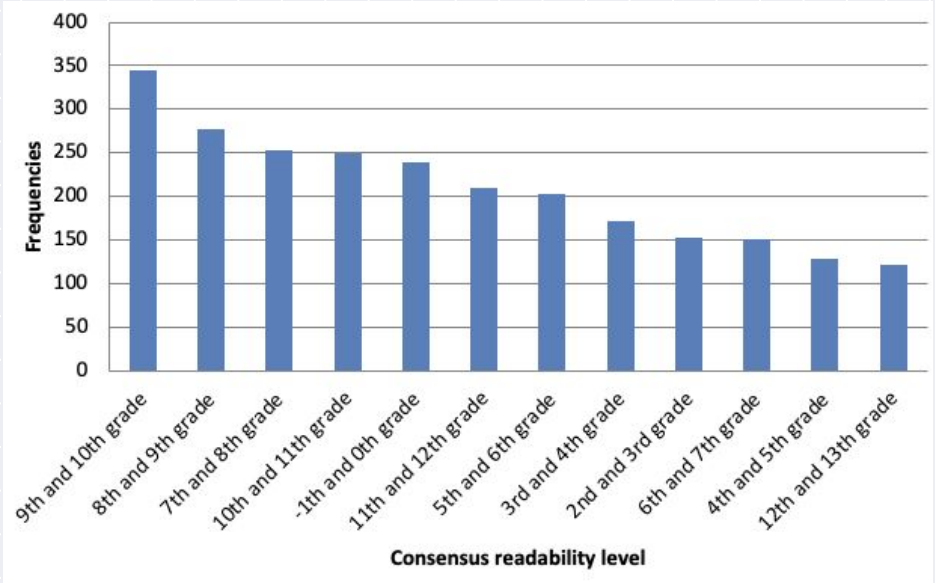


The Scarlet Letter: “Hester” and “Pearl”

Nominalization



Harry Potter Book 5



The Scarlet Letter