Independently Occurring Words and Book Popularity

Akansha Gupta[1], Chris Kolar[2]
Henry M. Gunn High School[1], Palo Alto Unified School District[2]

BACKGROUND AND SIGNIFICANCE

Readers and publishers are fascinated with trying to understand **what makes certain novels popular** and thus, more likely to be purchased. Several studies have examined book length while others have looked at the **frequency of specific words.** This project focuses on figuring out the effect of **Independently Occurring Words (IOW),** words which occur only once in the entire book.

Previous research on this topic

Publishers tell first-time authors to make sure they have about 80,000 words in their novels. However, most **books considered "classic" have a significantly higher word-count.** (~130,000 to the recommended ~80, 000).

In 2016, Matthew L. Jockers, associate professor of English at the University of Nebraska-Lincoln and Jodie Archer, a former acquisitions editor for Penguin UK created algorithm to predict a best-seller. Among other findings, the algorithm demonstrated that the **verb** 'need' is a much **stronger indicator of success** than the verb 'want.'

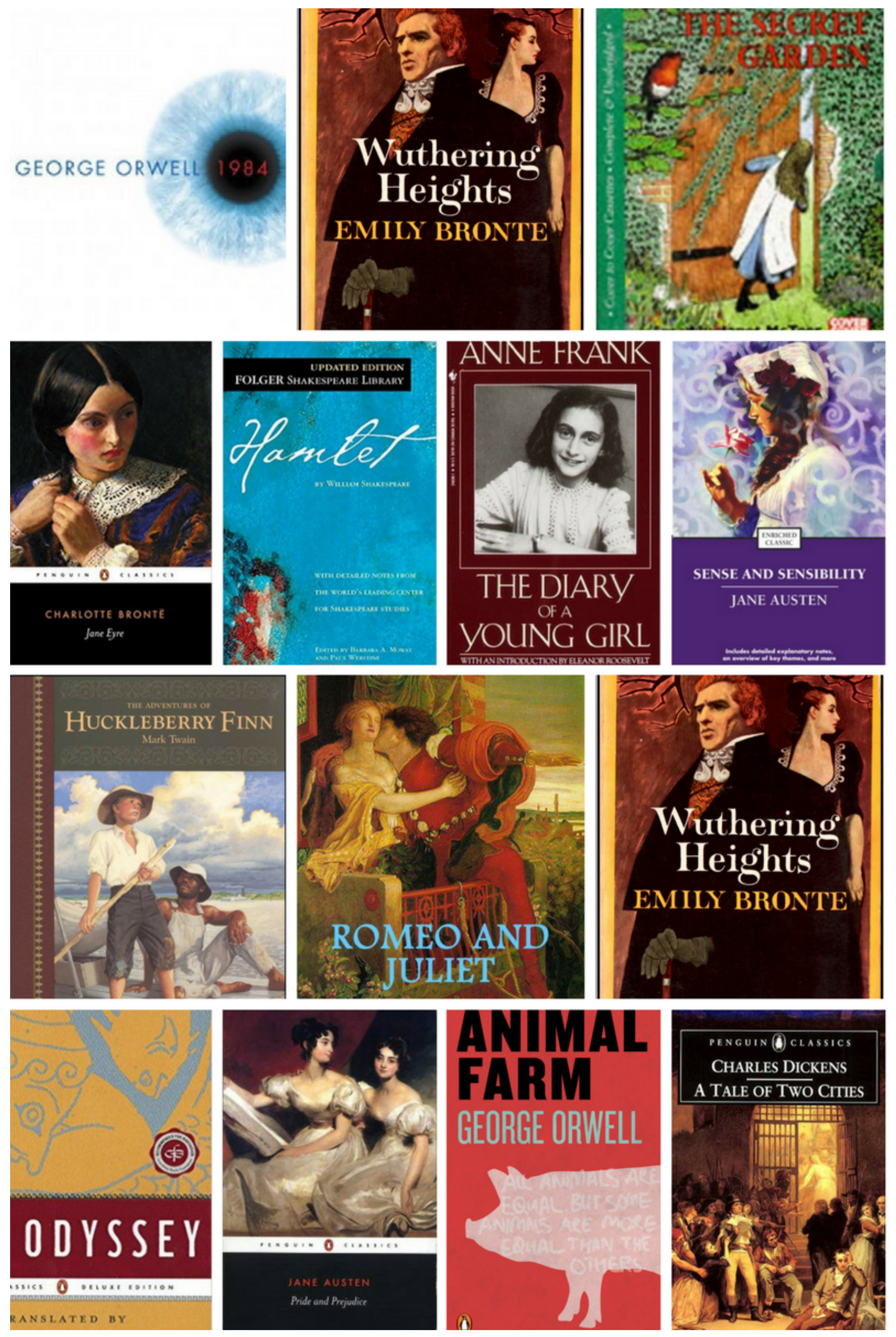
In individual works, such as *The Catcher in The Rye*, readers have tried to figure out the effect of the most repeated words. Blogger David Taylor noted that, unsurprisingly, the most frequent word used in *The Catcher in the Rye* is "goddamnit". Surprisingly, **the word which most people tend to recall** from the book (phony) is **not even in the top 40** words most used.

This project: IOW

The independently occurring words in a novel should give insights into **the repetition and vocabulary** an author needs to use to write a "popular book".

Words in different tenses or in adjective/adverb form will be clustered together. However, words sharing the same root will be considered as **separate words if the meaning of the two words is significantly different**. For example, "dictate", "dictator" and "dictating" should be clustered together and in the program, their frequencies added up. But, "dictionary" and "dictate" are recognized as different words.

RESEARCH METHODOLOGY



For this project, **popular books were downloaded** from Project Gutenberg, MIT-Stanford Literature project and other websites. In this project, a popular book is well defined and numerically verifiable: a book with over **500,000 ratings** as well as an average rating greater than **3.5/5 stars** on **Goodreads.com**

Based on this criterion, **14** books which fit these parameters have been chosen according to what was available in public domain.

Figure 1: Books used for analysis

These books were run through a word frequency counter and then a computer program created for this project. This computer program refers back to the University of Maryland's database, CatVar, which groups word clusters together. Thus, the program was able to **eliminate words which were used more than once in different forms**.

Figure 2: Word cloud of independently occurring words in the 14 books analyzed here (see Figure 1)

DATA ANALYSIS AND RESULTS:

	A Secret Garden	B 1984	C Huckleberry	D Tale Two Cities	E WutheringHeights	F BraveNewWorld	G SenseAndSensibility	H AnimalFarm	I Odyssey	J Hamlet	K RomeoJuliet	L DiaryYoungGirl	M DorianGray	N LittleWomen	Repeat Count
sale	■	■		■		■	■			■	■		■		8
ancestors	■	■		■					■		■		■	■	7
denied			■	■					■	■	■	■	■		7
everlasting			■	■		■	■	■		■		■			7
rightly		■		■	■	■	■			■		■			7
triumphant	■	■		■	■			■			■			■	7
untidy	■	■		■	■							■	■	■	7
wiping	■		■	■			■	■				■	■		7
asunder				■		■	■		■	■	■				6
awoke		■			■		■			■			■	■	6
beware			■	■					■			■	■	■	6
brim		■	■	■		■			■					■	6
burnt		■			■		■	■		■			■		6
deeply	■		■			■			■	■			■		6
deer			■	■		■				■		■	■		6

Chart indicates most common independently occurring words among the books analyzed. Of the 38 independently occurring words which show up in 6 books or more, 34 have **1 2, or 3 syllables.** 28 have 1 or 2 syllables. This indicates **correlation between relatively simple words and book popularity.**

Correlation coefficient between percentage of independently occurring words and rating of book is -3.33. To have negative correlation with 95% certainty the correlation coefficient needs to be -3.34. Thus **there is not a statistically significant correlation between popularity of a book and the percentage of independently occurring words**: words that appear in a book only once.

ACKNOWLEDGEMENTS AND REFERENCES:

Special thanks to mentor Chris Kolar for his advice and support.

1. "Great Novels and Word Count." Indefeasible. N.p., 2008. Web. 07 Oct. 2016.
2. Grey, Tobias. "An Algorithm to Predict a Bestseller." WSJ. Wsj.com, 31 Aug. 2016. Web. 14 Oct. 2016.
3. Taylor, David. "Word Cloud of The Catcher in the Rye." ~ Prooffreader.com. N.p., 2013. Web. 07 Oct. 2016.
4. Summers, Kate. "Adult Reading Habits and Preferences in Relation to Gender Differences." Reference & User Services Quarterly 52.3 (2013): n. pag. Web
5. "Text Analyzer." Text Analyzer - Text analysis Tool - Counts Frequencies of Words, Characters, Sentences and Syllables. N.p., n.d. Web. 26 Feb. 2017.
6. Habash, Nizar and Bonnie Dorr, A Categorial Variation Database for English, Proceedings of the North American Association for Computational Linguistics, Edmonton, Canada, pp. 96--102, 2003. [ ]
7. Orwell, George. "Animal Farm." Animal Farm, by George Orwell. N.p., n.d. Web. 26 Feb. 2017.
8. The Internet Classics Archive: 441 searchable works of classical literature. N.p., n.d. Web. 26 Feb. 2017.
9. "Hamlet: Entire Play." Hamlet: Entire Play. N.p., n.d. Web. 26 Feb. 2017.
10. "Nineteen eighty-four." Nineteen eighty-four. N.p., n.d. Web. 26 Feb. 2017.