Readability Analysis of Genres from the Brown Corpus

Felix Hammond

***Abstract*— This project analyzes the readability and syntactic complexity of different genres in the Brown Corpus. Using metrics like Flesch Reading Ease, Dale-Chall Score, and syntactic tree depth, we compare fiction and nonfiction texts. Statistical tests (ANOVA, t-tests) show significant differences, with nonfiction genres like "learned" and "government" being more complex and less readable than fiction genres such as "adventure" and "mystery." These results reflect how genre influences linguistic structure and readability.**

I. Introduction

Readability is a key component of effective written communication, influencing how easily a text can be understood by different audiences. This project investigates the readability of different written genres, specifically comparing fiction (e.g., mystery, romance) with nonfiction (e.g., government, academic writing). The central question is: Do nonfiction genres require higher reading levels and contain more complex sentence structures and vocabulary than fiction genres?

Understanding this difference has practical implications for educators, publishers, and developers of reading materials, especially when tailoring content to different audiences or education levels. Readability formulas such as the Flesch-Kincaid Grade, Dale-Chall Score, and average sentence length provide quantifiable measures for evaluating text difficulty.

II. Data

This project uses a corpus of English text excerpts from ten different genres. The dataset is categorized into the following genres:

Fiction Genres: Fiction, Mystery, Adventure, Romance, Humor Nonfiction Genres: Learned, Hobbies, Religion, Government, News Each genre contains multiple passages of varying length. The texts were pre-cleaned for basic formatting, but additional preprocessing was done (described below) to ensure consistency across readability calculations.

	Number of Words	Number of Sentences
learned	157984	7352
hobbies	70430	3991
government	60312	2647
religion	34109	1596
news	84650	4308
fiction	57930	4493
mystery	47802	4103
adventure	57949	4899
romance	58220	4744
humor	18057	1118

Fig. 1. Summary of data

III. Methods

Text data was preprocessed to remove extraneous characters, fix tokenization issues, and ensure accurate sentence and word segmentation. Special care was taken to filter out punctuation and numerals when calculating vocabulary statistics. The list of "easy" words used in the Dale-Chall Score was sourced from a local words.txt file derived from the official 3000-word list.

I used the following readability metrics:

- Flesch-Kincaid Grade: Approximates U.S. school grade level.
- Flesch Reading Ease: Higher scores indicate easier texts.
- Dale-Chall Readability Score: Focuses on unfamiliar (difficult) words based on a defined list.
- Average Sentence Length: A structural proxy for complexity.

The results of calculating each readability metric over the entire text for each genre are shown below.

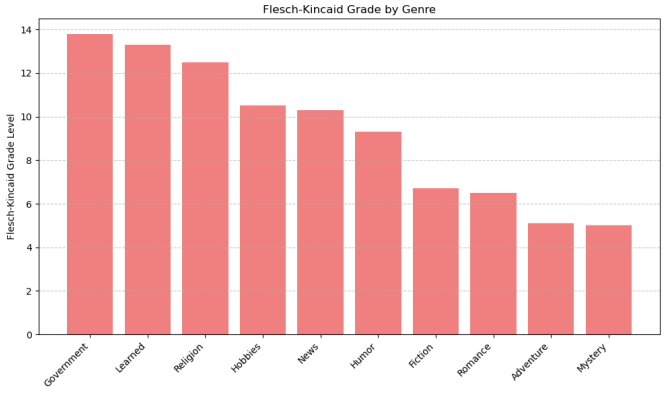

Fig. 2. Flesch Grade Level

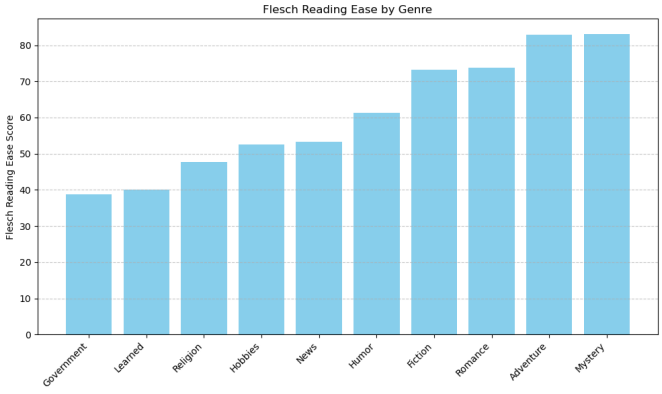

Fig. 3. Flesch Reading Ease

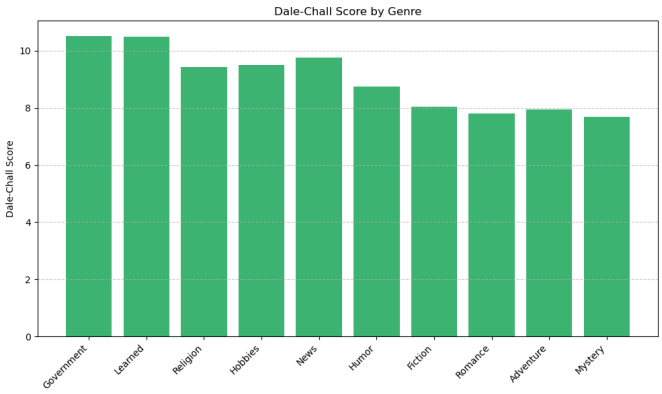

Fig. 4. Dale-Chall Score

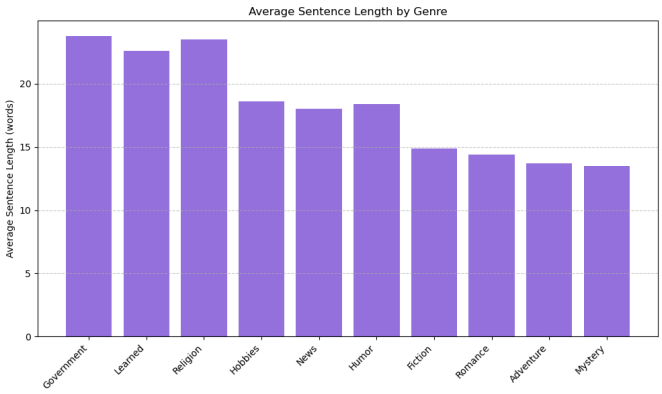

Fig. 5. Average Sentence Length

For analysis, I took multiple random samples from each genre and calculated all readability scores. Then a one-way ANOVA test was applied to determine if there are significant differences in readability between genres. I also performed a two-sample t-test comparing the fiction and nonfiction groups as defined above.

Hypotheses:

Null (H): There is no difference in readability scores across genres. Alternative (H): At least one genre has a significantly different readability score. For the t-test:

Null (H): There is no difference in readability scores between fiction and nonfiction. Alternative (H): Fiction and nonfiction differ in readability scores.

IV. Results

The ANOVA test showed a statistically significant difference in readability scores across genres ($p < 0.05$), leading us to reject the null hypothesis. The t-test comparing fiction and nonfiction also yielded a significant result ($p < 0.05$), confirming that nonfiction is, on average, harder to read than fiction. Learned and Government genres had the highest difficulty scores, while Adventure and Mystery had the lowest. Humor scored near the middle, which aligns with expectations — while often informal in tone, it may involve nuanced or idiomatic language.

V. Analysis

The results confirm a clear divide in readability between fiction and nonfiction genres. Fiction genres consistently scored lower on difficulty metrics, suggesting simpler language, shorter sentences, and more accessible vocabulary. This is consistent with their goal to entertain and appeal to broad audiences.

Nonfiction genres, especially Learned and Government texts, exhibited higher average sentence lengths and a higher percentage of difficult words. These genres often aim to inform or persuade, which may require precise and technical language, contributing to increased complexity.

Interestingly, Humor ranked in the middle. While often conversational, humor may include sarcasm, idioms, or cultural references that can introduce subtle complexity not captured purely by sentence length or word frequency.