Reading Time Analysis

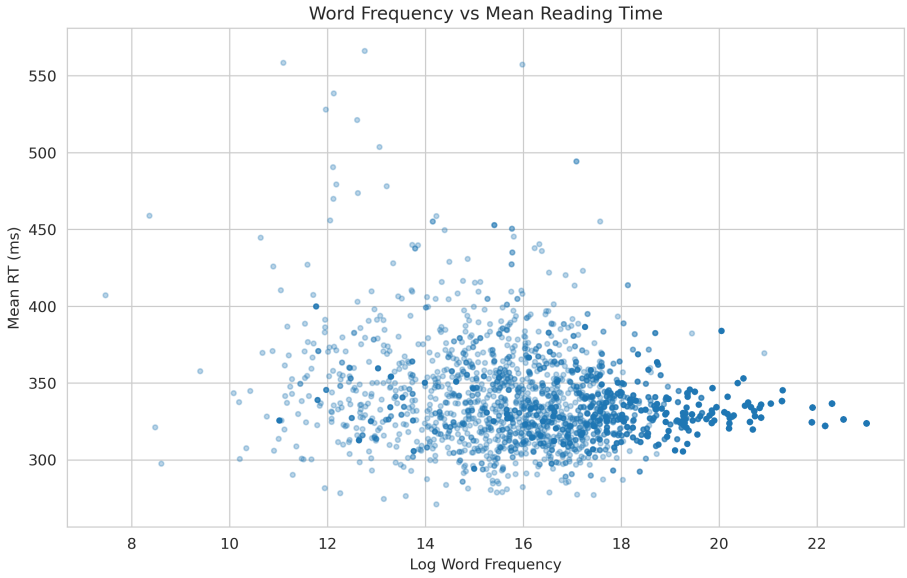
Name: Abhyudit Singh
Roll Number: 2023114009
Date: February 17, 2026
Code Repository: https://github.com/bitmap4/cp-a3

Dataset

Natural Stories corpus analysis with 848,875 reading time measurements across 8,301 unique words. Mean RT: 334.07 ms.

Part I: Correlations

Word Property Analysis



Correlations:

```
Length-RT:      r = 0.31, p < 0.001
Frequency-RT:   r = -0.25, p < 0.001
Length-Frequency: r = -0.71, p < 0.001
```

Length positively correlates with RT, frequency negatively correlates with RT. Strong length-frequency anticorrelation confirms Zipf's law.

Part II: Model Comparison

Frequency vs Predictability

Tested two regression models:

Model 1: RT ~ word_freq + word_length (R^2 = 0.0976, MSE = 453.68) **Model 2:** RT ~ -log(P_trigram) + word_length (R^2 = 0.1020, MSE = 451.47)

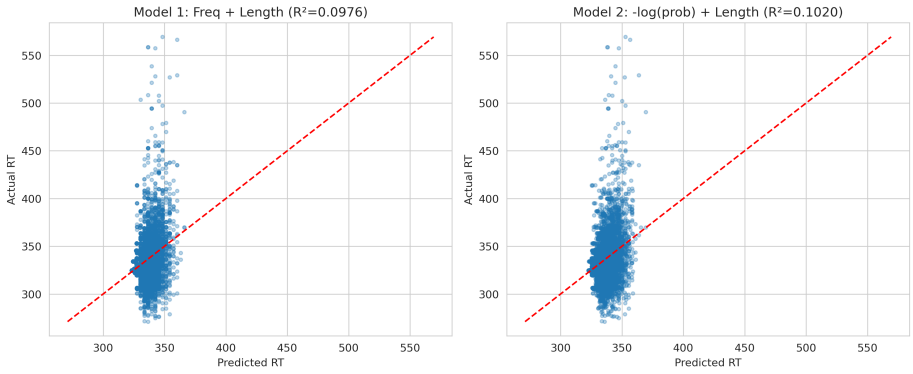


Figure 1: Model Battle Results

Model 2 outperforms Model 1. Contextual predictability explains more variance than raw frequency, supporting surprisal theory.

Content vs Function Words

Classified words by POS: content (NOUN, VERB, ADJ, ADV, n=4,883) vs function (ADP, DET, CONJ, PRON, PRT, n=3,347).

Content Words: - Model 1 (freq): R^2 = 0.0883 - Model 2 (context): R^2 = 0.0973 (better)

Function Words: - Model 1 (freq): R^2 = 0.1993 (better) - Model 2 (context): R^2 = 0.1158

Content words favor contextual predictability, function words favor frequency. Supports dual-route processing model.

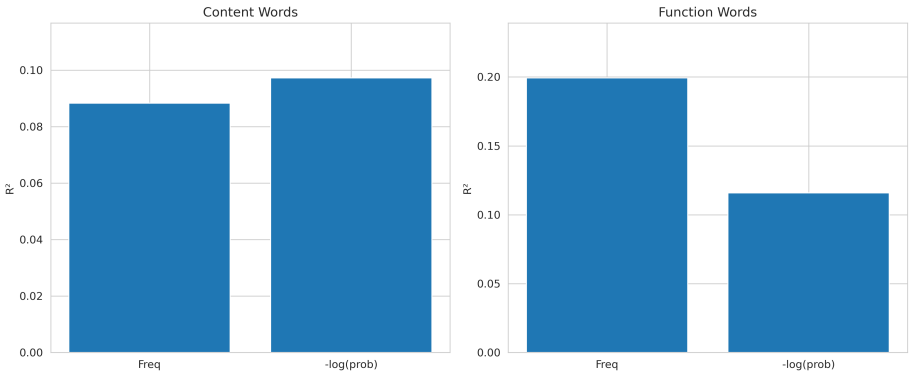


Figure 2: Word Type Processing

Part III: FOBS Model

Root vs Surface Frequency

Tested FOBS hypothesis using lemmatized roots:

Model 1: RT ~ surface_freq + word_length (R^2 = 0.0972) **Model 2:** RT ~ lemma_freq + lemma_length (R^2 = 0.0953)

Surface frequency marginally better. Limited evidence for morphological decomposition.

Pseudo-affix Analysis

Compared pseudo-affixed (finger, corner, butter, winter, number) vs real affixed (singer, owner, better, winner, hunter) words.

Results: - Pseudo-affixed: 323.12 ms - Real affixed: 325.73 ms

No pseudo-affix penalty observed. Morphological processing effects minimal in small sample.

Pseudo-affix Analysis

Compared pseudo-affixed (finger, corner, butter, winter, number) vs real affixed (singer, owner, better, winner, hunter) words.

Results: - Pseudo-affixed: 323.12 ms - Real affixed: 325.73 ms

No pseudo-affix penalty observed. Morphological processing effects minimal in small sample.

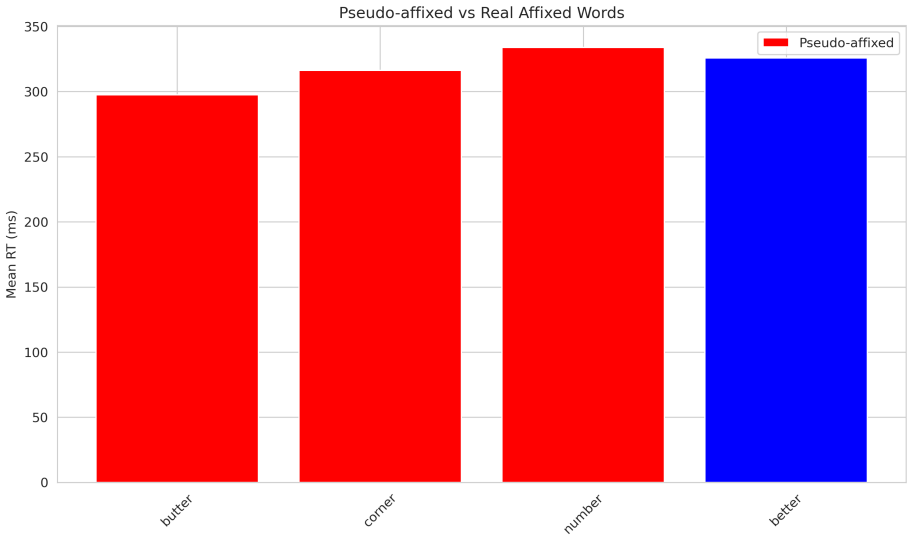


Figure 3: Morphology Test

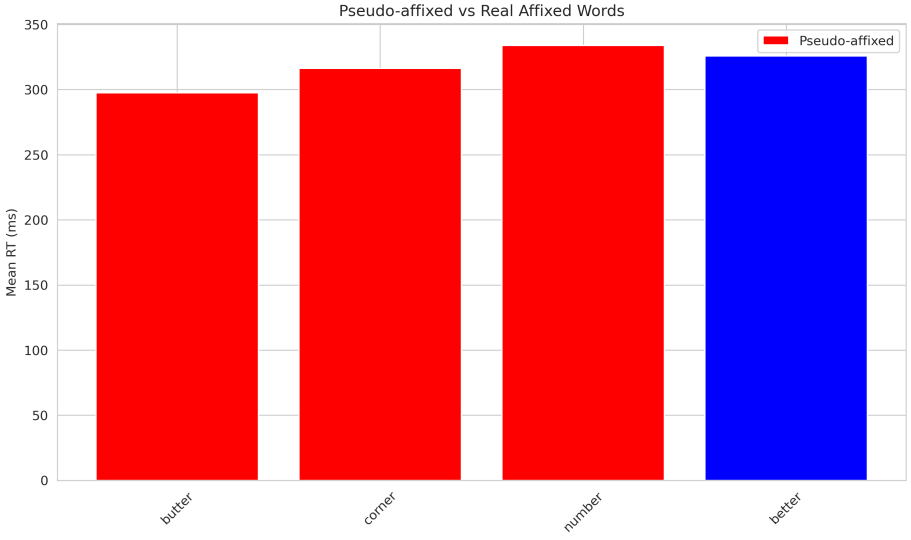


Figure 4: Morphology Test

Conclusions

Key Findings: - Length-RT r=0.31, Frequency-RT r=-0.25, Length-Frequency r=-0.71 - Contextual predictability outperforms frequency (R^2=0.102 vs 0.098) - Content words favor context, function words favor frequency - Limited morphological decomposition evidence

Implications: Reading involves parallel processing with context-sensitive lexical access. Frequency and length effects remain robust, while contextual expectations provide additional predictive power. Word class differences suggest specialized processing routes. Low R^2 values indicate significant unexplained variance requiring further investigation of sentence-level factors and individual differences.