COMPUTATIONAL PSYCHOLINGUISTICS

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

Project Report

Effects of Surprisal and Semantic Complexity on Reading Times

TEAM MEMBERS

Darur Lakshmipathi Balaji - 202114007
Revanth Gundam - 2021114018

May 8, 2024

Contents

1 Introduction **1**

2 Motivation **1**

3 Hypothesis **1**

4 Data Analysis **1**

5 Methodology **3**
5.1 Surprisal Analysis of Adverbs 3
5.2 Semantic Complexity Analysis 3

6 Results and Analysis **4**
6.1 Mean RT time Analysis 4
6.2 Entropy Reduction 5
6.2.1 Effect of Adverb Familiarity: 6
6.3 Surprisal vs. Reading Time 6
6.3.1 Analysis of Adverb Surprisal vs. Adverb Reading Time 6
6.3.2 Analysis of Adverb Surprisal vs. Sentence Reading Time 7

7 Conclusion **I**

8 References **I**

1 Introduction

This project, based in the field of computational psycholinguistics, investigates the effects of surprisal on the reading times of native Mandarin speakers. The study specifically explores how unexpected word choices and sentence structures impact reading efficiency and comprehension. This research leverages advanced computational methods to analyze linguistic data, aiming to contribute valuable insights to psycholinguistics, particularly in understanding the cognitive processes underlying language comprehension among L1 Mandarin speakers.

2 Motivation

Research in psycholinguistics often seeks to understand the cognitive processes involved in language comprehension. For non-native speakers, these processes can be significantly affected by linguistic features such as the complexity of sentence structure and the unpredictability of word usage. By studying these effects, we can develop more effective methods for language teaching and create linguistic models that better accommodate the learning needs of non-native speakers.

3 Hypothesis

The primary aim of this project is to quantify the impact of surprisal and semantic complexity on the reading times of L1 Mandarin speakers. The hypothesis is that higher surprisal and greater semantic complexity lead to longer reading times, reflecting increased cognitive load. The questions we are primarily looking into:

- Approximating the surprisal of adverbs
- Approximating the semantic complexity of a sentence

4 Data Analysis

We use a dataset which is a structured collection of data points designed to examine how surprisal and semantic complexity influence the reading times of L1 Mandarin speakers encountering English sentences. The data is meticulously organized to facilitate detailed analysis of adverb usage, its frequency, and the impact on reading times. Below is a breakdown of the data components and their significance in the project. The dataset consists of several key variables that provide comprehensive information for each participant and the sentences they read. Each entry includes the following fields:

- **Participant Information**: Each record includes demographic information such as the participant's ID, gender, and age, which allows for the assessment of demographic influences on the data.

- **Linguistic Features**:
  - Adverbs (Adv): Both common and rare adverbs used in sentences are recorded along with their frequency in spoken and written English. This categorization into 'common' and 'rare' is crucial for analyzing the impact of lexical familiarity on reading times.
  - Surprisal: Quantitative measures of surprisal are provided for each adverb, indicating the unexpectedness of encountering the adverb in the context of the sentence.
  - Word Length: The length of the adverbs used, which can be an indicator of lexical processing time.
- **Reading Times**:
  - Adverb Reading Time (Adv RT): Time taken by participants to read the adverb within the sentence.
  - Sentence Reading Time (Sentence RT): Overall time taken to read the entire sentence.
- **Parameters**:
  - Frequency: This denotes the frequency score of the adverbs as recorded in linguistic databases, which helps correlate commonality with reading times.
  - Entropy Measures: It includes calculated entropy values representing the complexity of the sentence structures. These are measured from GPT-2 probabilities.

This dataset provides detailed insights into how various elements of language complexity â from lexical rarity to structural unpredictability â influence the cognitive load and efficiency of language processing among non-native speakers. Using these datasets, the study aims to validate hypotheses concerning language comprehension challenges faced by Mandarin speakers when interacting with English, potentially guiding educational strategies and linguistic model improvements.

5 Methodology

5.1 Surprisal Analysis of Adverbs

Surprisal, a key concept in information theory applied to linguistics, measures the unexpectedness of an event, in this case, the use of an adverb in a sentence. Lower probability events carry higher surprisal and are hypothesized to increase reading times due to increased processing demands. The steps taken to calculate surprisal values were as follows:

1. **Tokenization and Adverb Extraction** Using the Hugging Face's GPT-2 large model, sentences were tokenized to map each word to an index in the modelâs vocabulary. Special attention was given to adverbs, which were extracted and tokenized separately.

2. **Sequence Preparation** Each sentence was prepared with Begin-of-Sentence (BOS) and End-of-Sentence (EOS) tokens to ensure that the probability estimations for each word were conditioned on realistic sentence contexts.

3. **Probability Estimation** The model computed the conditional logits (pre-softmax probabilities) for each token, focusing particularly on adverbs. These logits were converted into probabilities using the softmax function.

4. **Surprisal Calculation** Surprisal values were derived by taking the negative logarithm of the probability associated with each adverb. This quantifies how surprising an adverb is in the context of the sentence it appears in.

5. **Reading Time Comparison** Surprisal values were then correlated with empirical reading times measured from the participants, aiming to establish a direct relationship between the theoretical surprisal and the observed reading delays.

5.2 Semantic Complexity Analysis

Semantic complexity was analyzed through entropy reduction, which is based on the premise that each word in a sentence reduces uncertainty about subsequent words, thereby facilitating sentence processing. The process involved the following steps:

1. **Token Surprisal Calculation** Similar to the adverb analysis, the surprisal for each word in the sentence was calculated. This step was essential for accumulating the total surprisal across the sentence, which directly relates to its overall complexity.

2. **Entropy Reduction Computation** Sentential entropy was approximated by summing the surprisal values of each token within a sentence, from the first to the last. This cumulative measure provides an overall score of sentence complexity, assuming that higher total surprisal equates to greater semantic complexity.

3. **Complexity and Reading Time Analysis** The computed entropy values were then compared against reading times to assess how variations in semantic complexity influence language processing speeds in non-native speakers.

6 Results and Analysis

6.1 Mean RT time Analysis

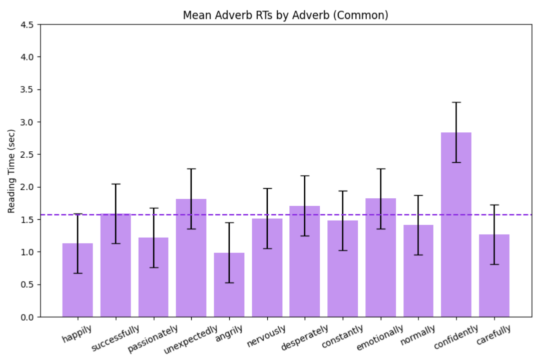

Figure 1: Mean RT for Adverb Common Condition

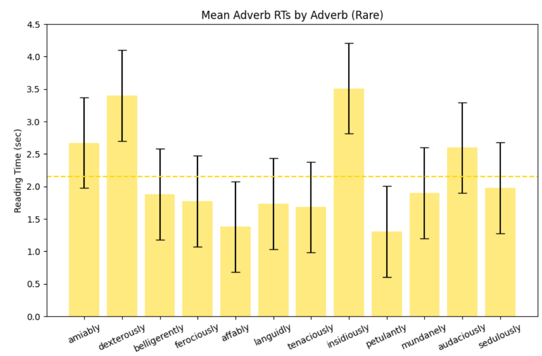

Figure 2: Mean RT for Adverb Rare Condition

The two graphs provided depict the mean reading times for different adverbs, categorized as either "common" or "rare". Each graph presents adverbs along the x-axis, with the corresponding mean reading times in seconds on the y-axis. These measurements are further detailed with error bars representing the variability or standard deviation of reading times for each adverb.

Rare adverbs generally show greater variability and longer mean reading times compared to common adverbs. This could imply that the familiarity with an adverb significantly affects reading speed and cognitive processing.

The graphs in Fig3,Fig4 represent the mean reading times of sentences containing either common or rare adverbs. These graphs are titled "Mean Sentence RTs by Adverb (Common)" and "Mean Sentence RTs by Adverb (Rare)", with each showing how different adverbs impact the time it takes to read sentences containing them.
Sentences containing rare adverbs tend to have longer reading times and greater variability. This is consistent with theories in psycholinguistics that suggest less frequent words require more cognitive resources to integrate into sentence structures.

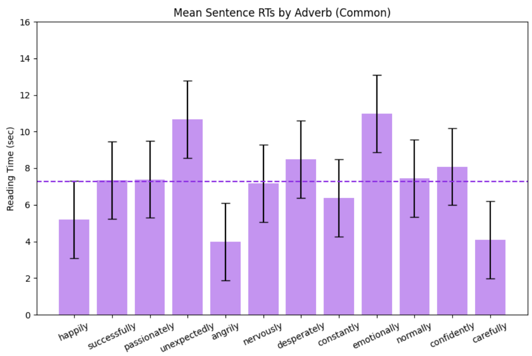

Figure 3: Mean RT for Sentence Common Condition

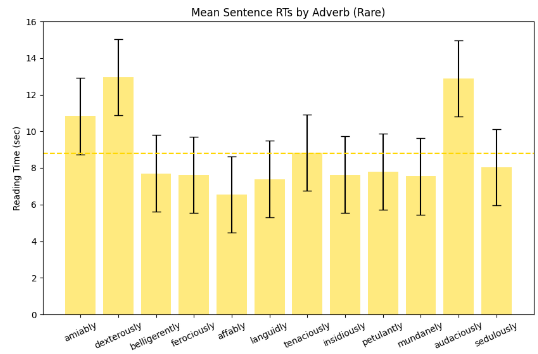

Figure 4: Mean RT for Sentence Rare Condition

6.2 Entropy Reduction

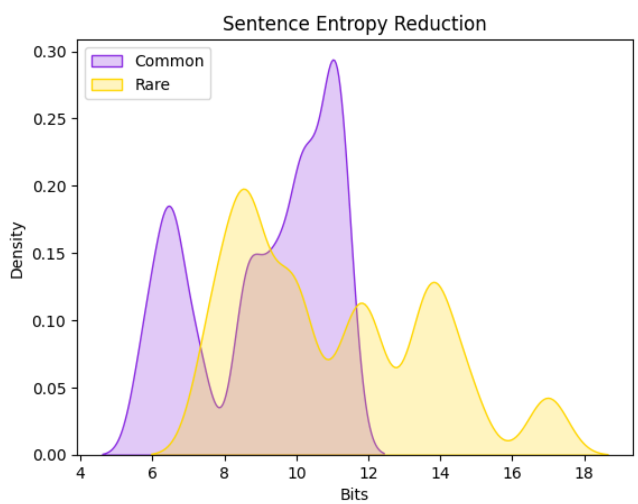

Figure 5: It presents the density of entropy reduction in sentences, comparing those with common adverbs to those with rare adverbs, plotted against the entropy values measured in bits.

1. **Common Verbs**: The distribution for common adverbs shows multiple peaks, with significant ones around 6 bits and 12 bits. This indicates that common adverbs contribute to entropy reduction at these specific levels, suggesting that they might generally be used in contexts that enhance predictability and clarity in sentence structure. The wider distribution implies that common adverbs contribute variably to reducing entropy, affecting sentence predictability in multiple ways depending on the context and usage.

2. **Rare Words**: The rare adverbs show a distribution with peaks around 8 bits and slightly above 14 bits. These peaks are less pronounced compared to those in the distribution of common adverbs, which may indicate that rare adverbs, while they do reduce entropy, do so less consistently or effectively. The overall lower density across the range suggests that rare adverbs have a less significant impact on reducing sentence entropy compared to common adverbs. This could be due to their less predictable usage or integration into typical sentence structures.

6.2.1 Effect of Adverb Familiarity:

The above graphs illustrates that familiarity with adverbs (as reflected by their commonality) plays a role in how much they can simplify sentence comprehension and predictability. Common adverbs, being more familiar, likely help readers process sentences with greater ease and certainty.

6.3 Surprisal vs. Reading Time

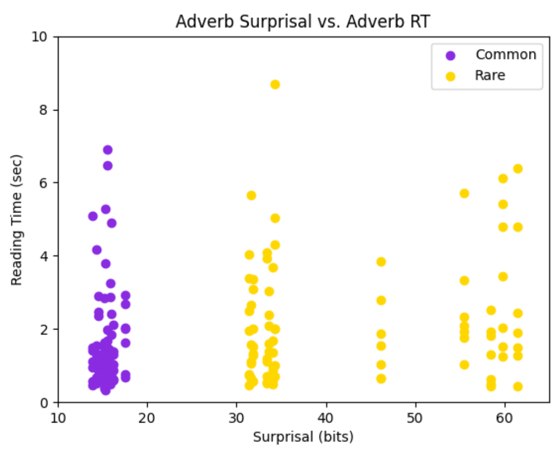

Figure 6: Surprisal vs Adverb Reading times

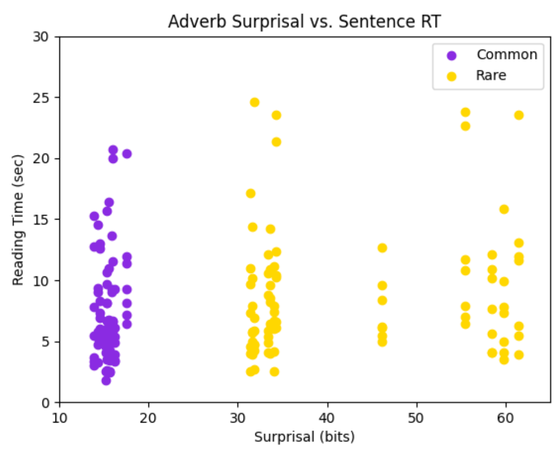

Figure 7: Surprisal vs Sentence Reading times

6.3.1 Analysis of Adverb Surprisal vs. Adverb Reading Time

In Fig.6 reading times for individual adverbs are plotted against their surprisal values.

1. **Common Adverbs:** These show lower surprisal values (mostly clustered between 10 and 30 bits) and generally lower reading times (mostly under 4 seconds). This suggests that common adverbs are more predictable and easier to process, resulting in faster reading times.

2. **Rare Adverbs:** These are plotted with higher surprisal values (ranging widely from 20 to over 50 bits) and show varied reading times, some of which are comparable to common adverbs,

while others are significantly longer. The increased reading time can be attributed to the higher surprisal and unfamiliarity associated with rare adverbs.

6.3.2 Analysis of Adverb Surprisal vs. Sentence Reading Time

In Fig. 7 the impact of adverb surprisal on the reading time of entire sentences is displayed which can be compared to the sentence complexity.

1. **Common Sentences:** Display a similar pattern to the adverb-level analysis, with lower surprisal values leading to generally consistent and lower sentence reading times, predominantly clustering around 5 to 10 seconds.

2. **Rare Sentences:** Reflect higher variability in both surprisal and reading times. Notably, some sentences with high surprisal adverbs also exhibit significantly extended reading times (up to 25 seconds), emphasizing the impact of adverb rarity and surprisal on the cognitive load and processing time for entire sentences.

The plots clearly demonstrate that higher surprisal values, which indicate less predictability and familiarity, correspond with increased reading times. This relationship is evident at both the adverb level and more pronouncedly at the sentence level, where the context surrounding the adverb further influences readability.

7 Conclusion

In this study, we investigated the impact of surprisal and semantic complexity on the reading times of L1 Mandarin speakers. Our hypothesis posited that higher surprisal and greater semantic complexity would lead to increased reading times due to a heightened cognitive load. The analysis and results have largely corroborated this hypothesis.

The data revealed that adverbs with higher surprisal values, especially those less familiar to readers, significantly increase reading times. This effect was evident in both the adverb-specific and full-sentence reading times. Specifically, rare adverbs showed longer reading times and greater variability compared to common adverbs, underscoring the role of lexical familiarity and surprisal in reading efficiency.

Moreover, our entropy reduction analysis demonstrated that sentences with higher semantic complexity (as quantified by total surprisal) necessitate longer processing times. This suggests that each additional bit of surprisal in a sentence's structure compounds the overall cognitive effort required for comprehension.

Overall, the findings validate the initial hypothesis and highlight the importance of surprisal and semantic complexity in shaping the reading experiences of non-native speakers. These insights can inform future psycholinguistic research and contribute to the development of more effective language teaching strategies and computational models that accommodate the processing needs of L1 Mandarin speakers. This study not only advances our understanding of cognitive processes in language comprehension but also sets the stage for further explorations into the intricate dynamics between language structure and reader interaction.

8 References

- Language Models are Unsupervised Multitask Learners
- https://huggingface.co/docs/transformers/model-doc/gpt2
- Hale J. Uncertainty about the rest of the sentence. Cogn Sci. 2006 Jul 8;30(4):643-72. doi: 10.1207/s15516709cog0000-64. PMID: 21702829.