096222: Language, Computation and Cognition
Surprisal and RTs

Daniel Shemesh (211388251) and Eylon Efraim (207427428)
Technion – Israel Institute of Technology
daniel.sh@campus.technion.ac.il
ef@campus.technion.ac.il
Project Github Repository

Abstract

This report presents an investigation into the relationship between surprisal estimates derived from n-gram and Recurrent Neural Network (RNN) language models, and their correlation with human reading times. The paper expands on the work of (Smith and Levy, 2013) by training an RNN language model on the Penn Treebank dataset and comparing its surprisal estimates with those of the n-gram model. Additionally, the report partially confirms and expends upon the analysis presented in (De Varda and Marelli, 2022), demonstrating the RT-suprisal relationship across various languages is present even when using bi-gram language models with limited contextual information.

1 Introduction

1.1 Background and Motivation

Understanding the relationship between word predictability and reading time is essential for studying natural language comprehension. In their paper, (Smith and Levy, 2013) investigated this relationship using probabilistic language models and eye-tracking data. They discovered a logarithmic effect, wherein reading time decreases by a constant amount for each unit increase in the logarithm of word probability. This finding has significant implications for our understanding of reading behavior and motivates further exploration into the effectiveness of different language models in predicting reading times.

Surprisal is a measure of how unexpected a word is in a given context. It is defined as the negative logarithm of the word probability given the preceding words:

$$S(w_i) = -\log P(w_i|w_1, ..., w_{i-1})$$

where $S(w_i)$ is the surprisal of word w_i, $P(w_i|w_1, ..., w_{i-1})$ is the probability of word w_i given the preceding words $w_1, ..., w_{i-1}$, and log is the natural logarithm.

1.2 Objectives

Building upon the work of (Smith and Levy, 2013), our project aims to extend the investigation of the relationship between word predictability and reading time in two key ways. Firstly, we train a Recurrent Neural Network (RNN) language model on the Penn Treebank dataset to obtain surprisal estimates. Comparing these estimates with the n-gram model used by (Smith and Levy, 2013) will allow us to examine the effectiveness of different language models in predicting reading behavior.

Secondly, we expand upon that analysis by comparing the results obtained with the simple language models to ones obtained with General Additive Models, as well as results obtained with self-paced corpora instead of eye-tracking corpora.

Thirdly, we attempt to confirm the results of (De Varda and Marelli, 2022) with a simple bigram model, with the purpose of investigating whether the limited context provided by a single word suffices for the purposes of predictive processing, such that the following words will be easier to read.

2 Structured Tasks

2.1 Training the RNN language model

The Penn Treebank dataset (Marcus et al., 1993) is a corpus of English text that has been annotated with syntactic and semantic information. We trained a recurrent neural network (RNN) for 40 training epochs on the training set of this dataset, using this Jupyter notebook here. Then, we used the trained RNN model to predict the surprisals on the validation set of the same dataset.

We harmonized the results of the RNN and the N-gram and compared the predicted surprisals of the two models. Presented below are the results.

Language Model	Correlation to RTs
RNN surprisals	0.2020939454327774
n-gram surprisals	0.19996728191611632

Table 1: Correlation of surprisals to Reading Time

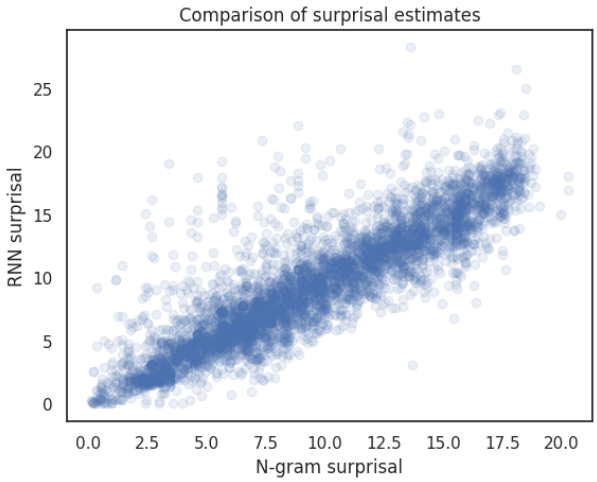

Figure 1: Comparison of surprisal estimates between RNN and N-gram models

2.2 Comparison of RNN and N-gram models

2.2.1 Question 1 - Correlation between surprisal estimates and reading times

As shown in Table 1, the RNN model has a slightly better correlation with human reading times, 0.202, compared to the n-gram model's 0.199, but the difference is very slight.

2.2.2 Question 2 - Surprisal estimates visualization

The graph (displayed in Fig. 1) shows a very high correlation between the n-gram surprisal values and the RNN model's surprisal values. We can conclude that the models are indeed well-matched, although there seem to be more outliers among words for which the n-gram model estimated a low surprisal value and the RNN model a higher one. Based on the graph, the RNN model seems to assign greater surprisal values than the n-gram model. Fig. 2 illustrates this difference.

Unlike our estimate based on Fig. 1, there does not appear to be a clear trend of lower RNN values compared to n-gram values, however there do appear to be some extreme outliers for whom the RNN values is much greater, which is probably the cause of our false initial perception.

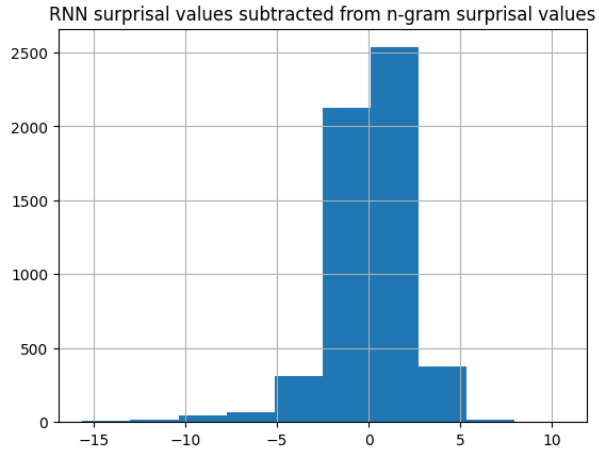

Figure 2: RNN surprisal values subtracted from n-gram surprisal values

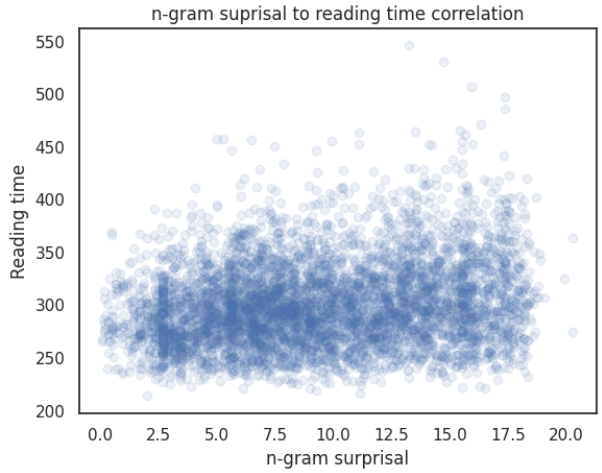

Figure 3: n-gram surprisal to reading time correlation

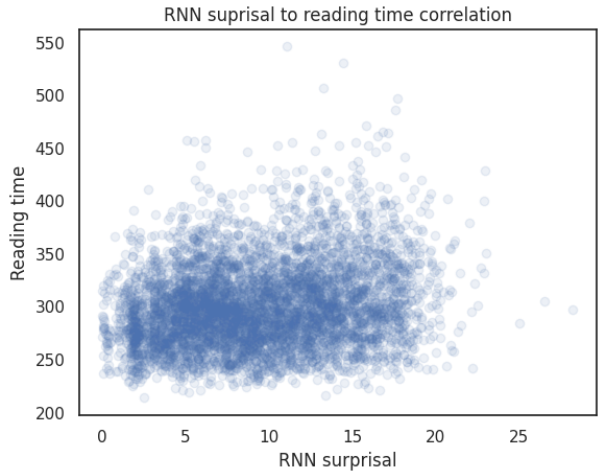

Figure 4: RNN surprisal to reading time correlation

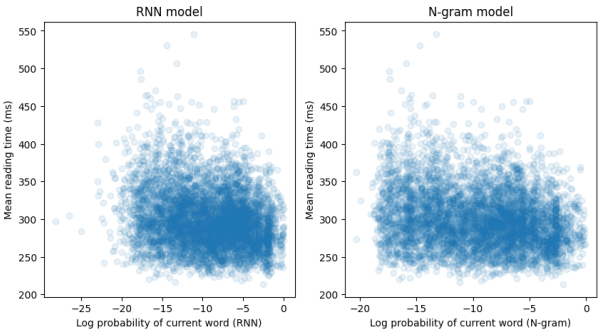

Figure 5: mean reading time against the log probability of the current word

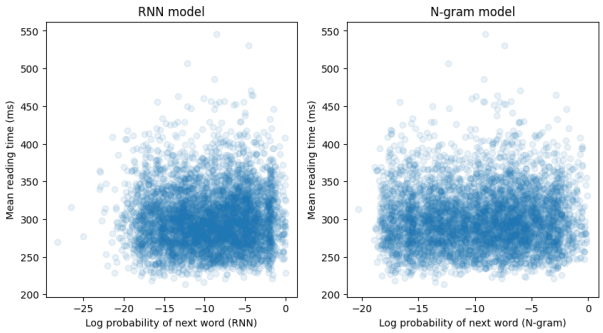

Figure 6: mean reading time against the log probability of the next word

2.2.3 Question 3 - Examination of sentences with discrepant surprisal estimates

We'll check the extreme outliers found in Fig. 3 and Fig. 4:

Points of Interest

Token John

Sentence She was the John one of those Atlantic that had sailed to to bring and bullets to the U.S.

Token I

Sentence Down in New was a flier in the right place at the right Robert S. a native New had been a World War I flying and one of the original planners of the Concord

Note that the sentences appear without unique words that were replaced by a token.

2.2.4 Question 4 - Analysis of spillover effect

As shown in Fig. 5 and 6, the ratio between the log-probability and the mean reading time appears to be very similar between the current word and the next word's reading time, which would indicate that the spillover effect applies. The effect exists for both models, but the shape of the ratio's distribution is different, with the log-probabilities of the n-gram model having a greater variance than the RNN's while having similar reading times.

3 Semi-Structured Tasks

3.1 Fitting and plotting RT surprisal curve using GAM

We plotted the RT surprisal curve using a General Additive Model (GAM). The model includes control variables for log-frequency and word length.

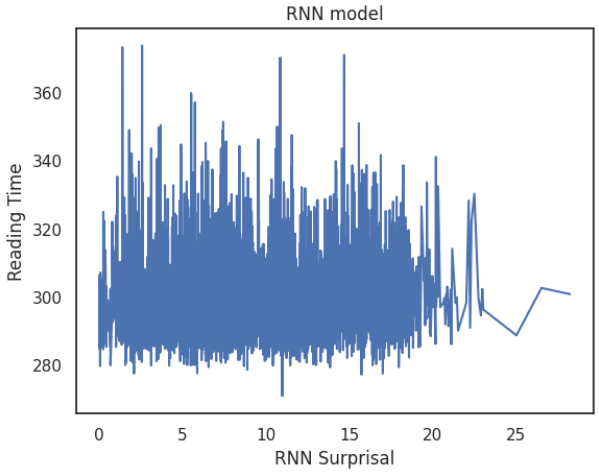

Figure 7: GAM Model Plot: Current Word RNN Surprisal vs. Reading Time

We examine both current word and spillover effects.

The results are presented in Fig. 7 and 8, which present the relationship between the current word's RNN-based surprisal and the reading time when using a General Additive Model, as well as in Fig. 9 and 10, which present that same relationship with the previous word's surprisal.

3.2 Surprisal analysis with a different reading times corpus (self-paced reading)

The Natural Stories Corpus (Futrell et al., 2021) is a collection of English stories, containing self-paced reading time data. We used it as an alternative RTs corpus, and once again we trained a RNN model. The results are shown at Fig. 11, Fig. 12, Fig. 13 and Fig. 14.

As seen in Table. 2 the correlation between the surprisal values and the reading time is 0.0152, and is lower than the 0.0202 correlation obtained using the Penn Treebank corpus.

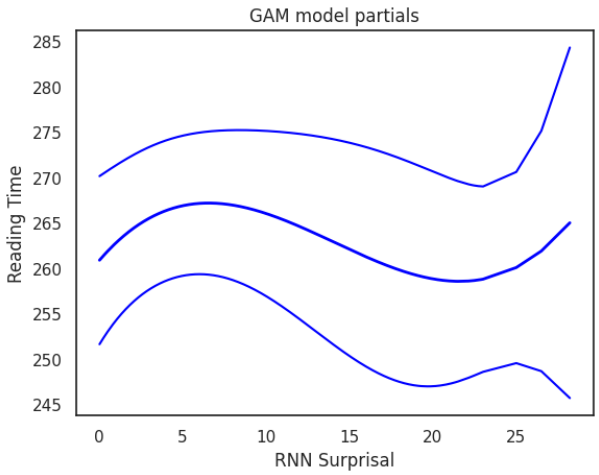

Figure 8: GAM Model Partial Plot: Current Word RNN Surprisal vs. Reading Time

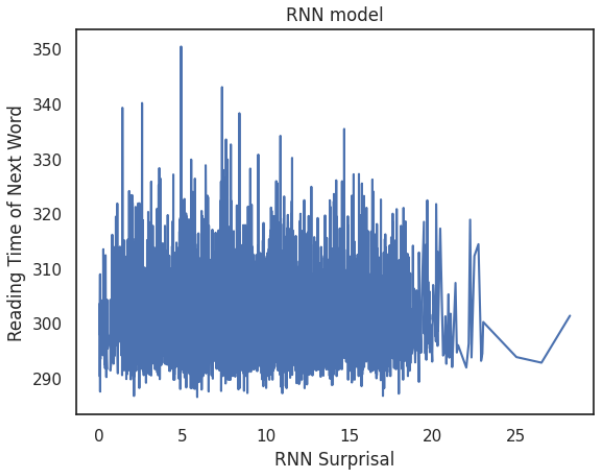

Figure 9: GAM Model Plot: Next Word RNN Surprisal vs. Reading Time

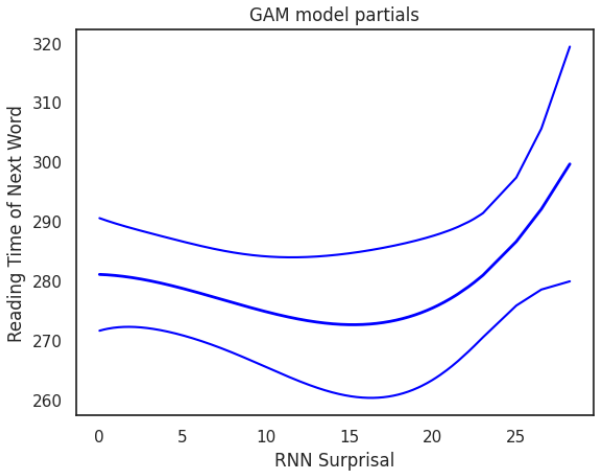

Figure 10: GAM Model Partial Plot: Next Word RNN Surprisal vs. Reading Time

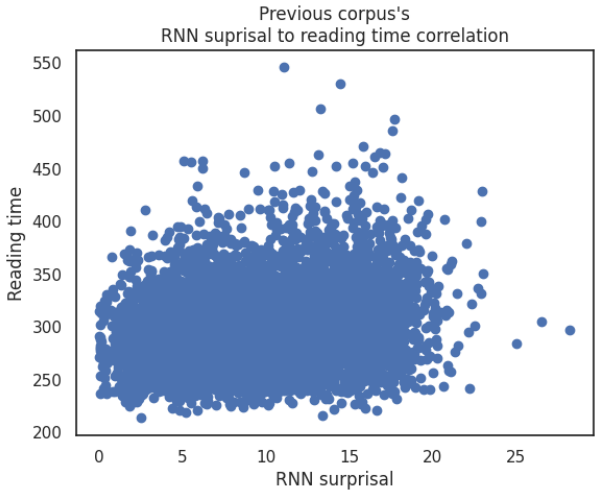

Figure 11: Penn Treebank corpus's RNN surprisal to reading time correlation

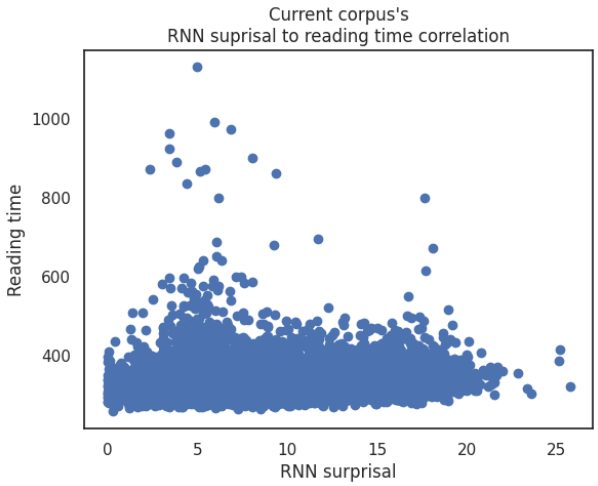

Figure 12: Current corpus's RNN surprisal to reading time correlation

Corpus	Surp. Correlation to RTs
Penn Treebank	0.2020939454327774
Self-Paced	0.015120480117041299

Table 2: Correlation of surprisals to Reading Time

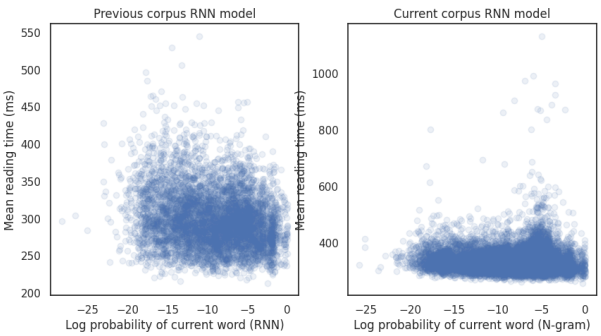

Figure 13: Comparison of corpora on Mean RT by log probability of current word (RNN)

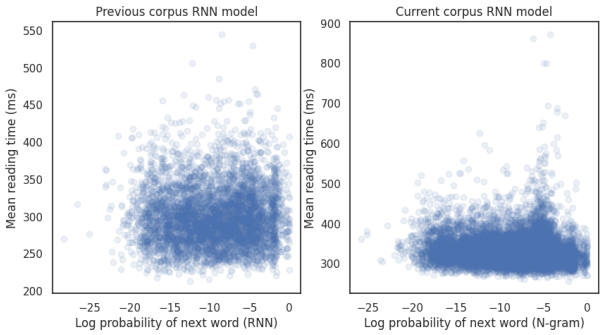

Figure 14: Comparison of corpora on Mean RT by log probability of next word (RNN)

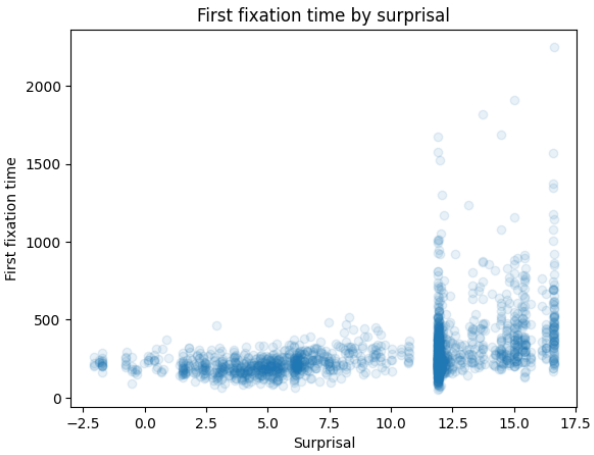

Figure 15: First fixation time by surprisal - German

4 Open-Ended Tasks: Comparing the Effects of Surprisals among various Germanic languages

4.1 Introduction

Building upon the findings of (De Varda and Marelli, 2022), we will run a cross-lingual comparison of bigram models on Germanic languages.

4.2 Training Corpora

The Leipzig Corpora (Goldhahn et al., 2012) is a collection of or large-scale language datasets, compiled by the Natural Language Processing Group at Leipzig University. We used Leipzig's datasets in Dutch, English and German as training data.

4.3 Eye-tracking Data

GECO (Cop et al., 2017) is a corpus of eyetracking data from monolingual and bilingual readers of a novel. PoTeC (Jäger et al., 2021) is a corpus of eyetracking data from experts and non-experts reading scientific texts in German.

4.4 Methods

4.4.1 Bigram model

We used a bigram model to train and evaluate the language data for each specific language. Our motivation stemmed from the fact that the original paper utilized a more intricate language model, namely mBERT. We sought to determine if similar results could be achieved by using a simpler language model.

Consequently, we computed the surprisal estimates provided by the bigram model on the respective eye-tracking dataset for each language. The results are presented below.

4.4.2 Results

The numerical results are presented in Table 3 and Table 4.
We fitted six linear regression models: 2 for each language, analysing the relation between the first fixation duration and surprisal, as well as the relation between the total reading time and surprisal. Using these models, we've predicted surprisal values for each duration value, and used Welch's unequal t-test (as the variances of the two groups are quite different) to measure the statistical significance of the results. We can observe that in both the English and Dutch languages, the relationship between surprisal and reading times (both total and first fixations) is clearly significant, while in the German language the relationship is not statistically significant.

As (De Varda and Marelli, 2022) notes, the first fixation duration is considered a sign of predictive processing, and as such, the presence of a statistically significant corellation between it and the surprisal values in several langauges is particularly indicative of the existence of a relation between surprisal values and human processing time.

As can be expected, we can observe that the total reading time in all three langauges is generally greater than the first fixation time, but it is interesting to note that the difference appears more pronounced in German, a language charecterized by many long, compound words. The graphs illustrate our numerical results, demonstrating an easily observable correlation between reading duration and surprisal for both English and Dutch, and a much less clear and significant difference in German.

Language	First fixation duration						
	ρ	σ^2	SE	β	$\hat{\sigma}^2$	t	pval
Dutch	0.39	3632.49	0.61	0.02	2.99	5.15	2.65e-7
English	0.33	4315.14	0.73	0.02	1.97	3.75	0.00
German	0.40	34747.99	4.28	0.00	3.38	0.71	0.47

Table 3: Statistics of first fixation duration by language. ρ indicates the correlation coefficient between duration and surprisal, σ^2 indicates the variance of the duration, β the linear regression coefficient between the duration and surprisal values. $\hat{sigma}^2$ indicates the variance of the regression model's predictions, t indicates the t-statistic of Welch's unequal-variance t-test, and $pval$ indicates the p-value of that same test.

Language	Total reading time						
	ρ	σ^2	SE	β	$\hat{\sigma}^2$	t	pval
Dutch	0.40	12037.76	1.11	0.01	3.21	7.13	1.36e-12
English	0.39	19144.05	1.54	0.01	2.80	6.38	1.99e-10
German	0.39	207798.40	10.47	0.00	4.43	0.24	0.80

Table 4: Statistics of total reading time by language. ρ indicates the correlation coefficient between reading time and surprisal, σ^2 indicates the variance of the reading time, β the linear regression coefficient between the time and surprisal values. $\hat{sigma}^2$ indicates the variance of the regression model's predictions, t indicates the t-statistic of Welch's unequal-variance t-test, and $pval$ indicates the p-value of that same test.

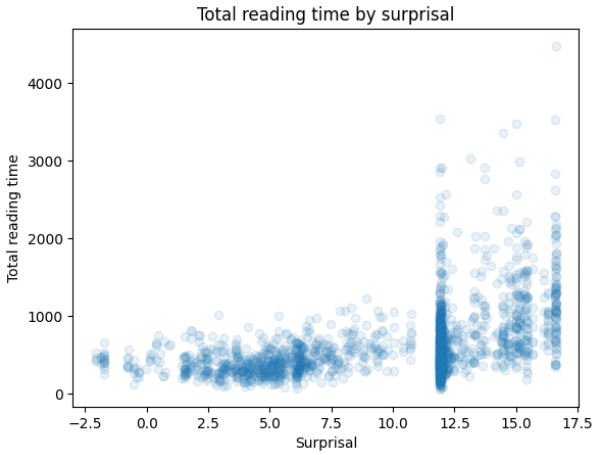

Figure 16: Total reading time by surprisal - German

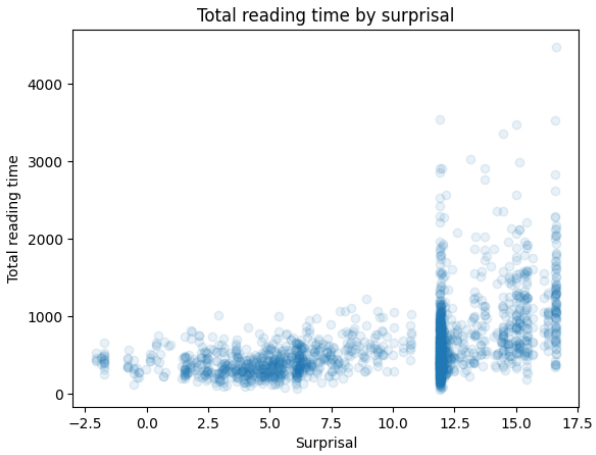

Figure 17: Total reading time by surprisal - German

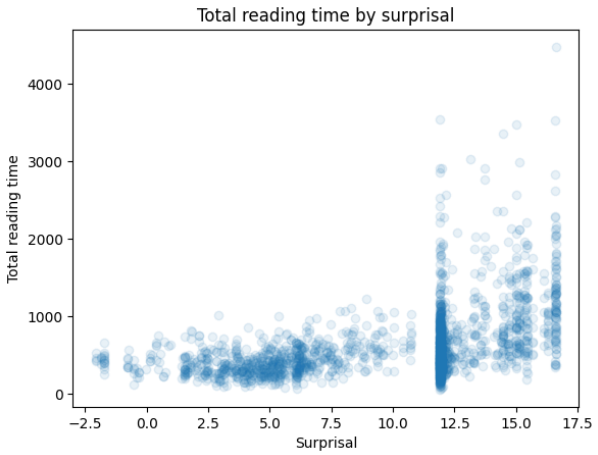

Figure 18: Total reading time by surprisal - German

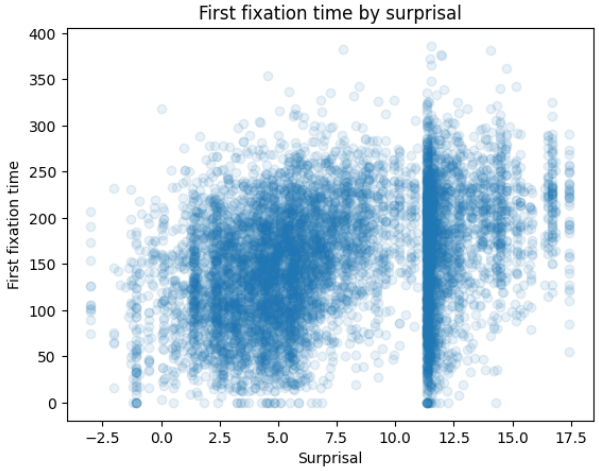

Figure 19: Total reading time by surprisal - German

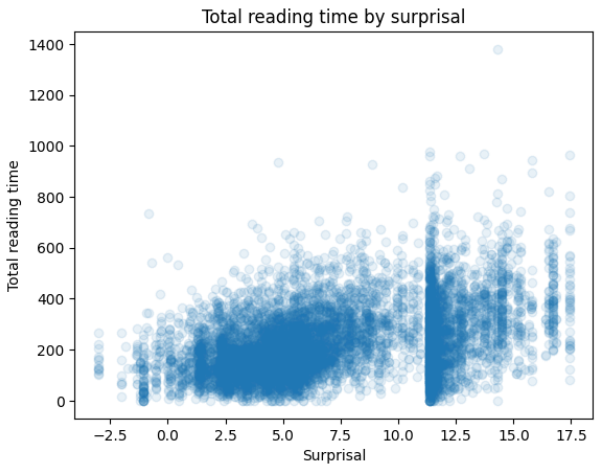

Figure 20: Total reading time by surprisal - German

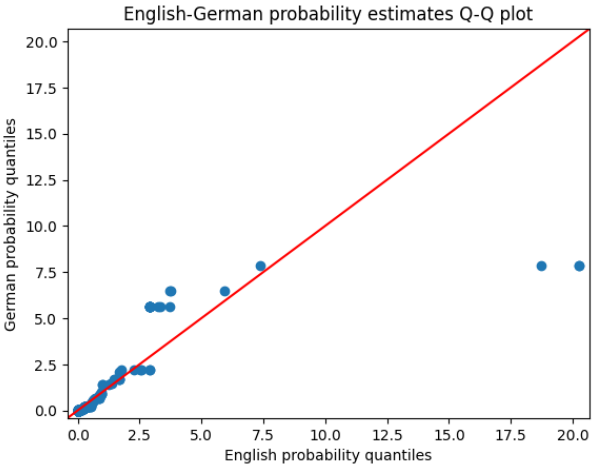

Figure 23: Total reading time by surprisal - German

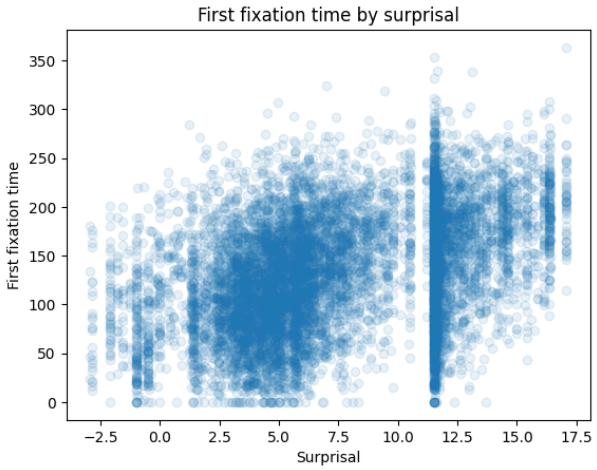

Figure 21: Total reading time by surprisal - German

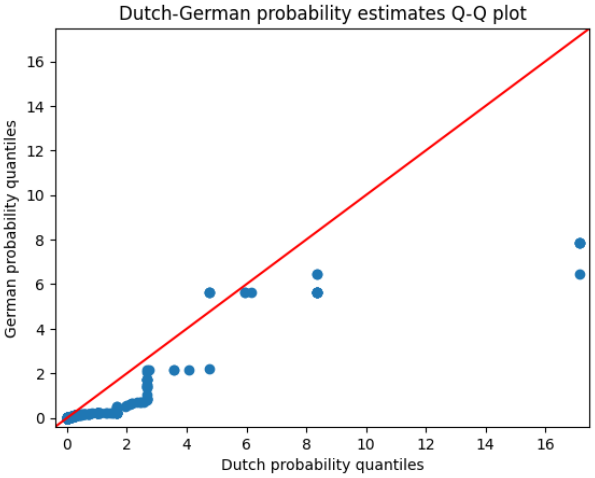

Figure 24: Total reading time by surprisal - German

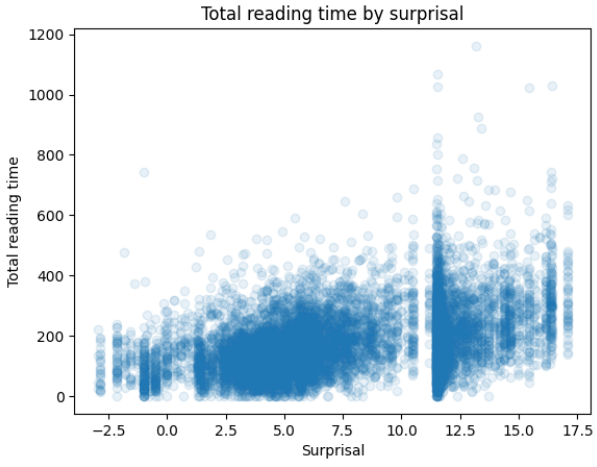

Figure 22: Total reading time by surprisal - German

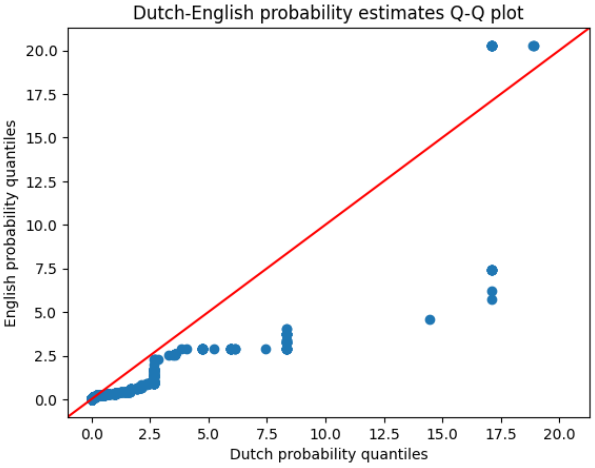

Figure 25: Total reading time by surprisal - German

4.5 Discussion

We manage to confirm that in several languages, the ability to predict future words is a significant predictor of the time taken to, and thus difficulty involved with, reading them. The existence of this relation even when the surprisal is calculated using an exceedingly simple bi-gram model is noteworthy, as it suggests that a single previous word's worth of context is sufficient for the brain's predictive processing to anticipate the next word to a such an extent that the difficulty in reading it would be significantly affected.

English and Dutch both count among those languages for which (De Varda and Marelli, 2022) found a significant relation between surprisal and reading times, and our project confirms those results for the case of a simple bi-gram language model.

However, unlike (De Varda and Marelli, 2022), we've found no statistically significant relation between reading times and surprisal in the German language. We believe this is the result of a limitation present in both our model and that used in the original research paper, which was handled in two different ways: The German language is notable for its many compound nouns, many of which connect three or more nouns. The mBERT model used by (Marelli and de Varda, 2022) can handle such compound nouns gracefully by dividing the compounded word into multiple tokens, but during the study's analysis, all multi-token word were dropped. As such, this unique and challanging trait of the language did not come into account. On the other hand, our bi-gram model and subsequent analysis did consider such compound words, but did not posses the capacity to seperate them into multiple tokens, and as such included many more unique and near-unique words than would be present in similar texts in comparable languages. Further analysis would be necessary to make proper conclusions, but we believe this seperate approach to compound words is the root of the differnce in results w.r.t. the German language.

References

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49:602–615.

Andrea De Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Lena Jäger, Thomas Kern, and Patrick Haller. 2021. Potsdam textbook corpus (potec): Eye tracking data from experts and non-experts reading scientific texts. *OSF*, 10.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.