Exploring Surprisal from Various Language Models for Predicting English Reading Times of People with Different Language Backgrounds

by

Shannon Clark

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

© Shannon Clark, 2023

Abstract

Surprisal estimated by language models is predictive of reading time in first-language (L1) reading. Research is emerging to determine whether this observation extends to reading in a second language (L2). Current attempts to characterize differences in the predictive power of surprisal for L1 and L2 reading times lack exploration of the reader's language background. As such, this thesis aims to evaluate the performance of surprisals derived from various language models for predicting English reading times of people with different L1s. To this end, we trained nine language models that varied in the extent of syntactic information, lexical information, and preceding context they used to compute surprisal. Next, we developed generalized additive mixed models to predict the English reading times of L1 speakers of English, Chinese, Korean, and Spanish using surprisal. Our results showed several commonalities. First, the best-performing surprisal for all language backgrounds was derived from a standard n-gram or an n-gram with added part-of-speech tags. Second, the lexical portion of total surprisal from a probabilistic context-free grammar performed more poorly than the syntactic portion. Last, out of the surprisals estimated using only syntactic information, those that accounted for the hierarchical structure of sentences outperformed the one based purely on sequential representations. Apart from these similarities, we observed differences by language background. It appears that surprisal computed using richer context performed better for L1 speakers of left-branching languages. It also seems that surprisals derived using hierarchical syntactic information performed better for languages with a different word order than English. Further research is needed to fully characterize these differences in performance in terms of the linguistic features of the reader's L1 and the way each language model computes surprisal. Our work shows that a variety of language models produce surprisals predictive of L1 and L2 reading times in English. Since the performance of these surprisals varied by the reader's L1, our work suggests that it is important to consider language background when using language models in the study of L2 reading.

Preface

The online program used to collect the reading time data used in this thesis was made by Max Helfrich. He collected the data with approval from the University of Pittsburgh's Institutional Review Board under the protocol name "Behavioral Studies of Reading and Language" (STUDY1904037). This project was sponsored by the National Science Foundation of the U.S. (PI: Lin Chen; Co-PI: Charles Perfetti, BCS-2118195). The data collected from this project will be made available to the public in the future. I have detailed the methods of data collection in Section 4.1.

The project detailed in this thesis was initially started by Daniela Teodorescu who did work training the n-gram and probabilistic context-free grammar models before handing it off to me. As the project progressed, I made changes to the training data and retrained these models.

Aside from the reading time data collection, I was responsible for all other aspects of the project including training the language models, developing the generalized additive mixed models, performing the statistical analyses, and writing this thesis.

The code used for this thesis can be found at https://github.com/EdTeKLA/IncrementalReadingLanguageModelling. This repository contains both original code and code that I adapted from the work of others. This distinction is marked clearly in the repository with reference to the code's original source.

Acknowledgements

I would like to thank my supervisor, Carrie Demmans Epp, for her support and encouragement. Her expertise guided me through the research process and her insightful feedback improved my work immensely.

I would also like to express my gratitude to Charles Perfetti, Lin Chen, Gaisha Oralova, Alona Fyshe, and Daniela Teodorescu for their valuable input and helpful discussions throughout this project.

Finally, I am so grateful to my family and friends who have always believed in me. To my parents and siblings, thank you for your endless love and support. To John, Aysha, Claire, and Annie, thanks for all the laughter and fun that brought so much joy to this journey.

Table of Contents

List of Tables

List of Figures

Abbreviations

AIC Akaike information criteria.

GAMM Generalized additive mixed model.

L1 First language.

L2 Second language.

PCFG Probabilistic context-free grammar.

POS Part-of-speech.

RNN Recurrent neural network.

RNNG Recurrent neural network grammar.

SOV Subject-object-verb.

SVO Subject-verb-object.

Glossary of Terms

Alphabetic language A language that uses a writing system where symbols map to the smallest units of sound called phones.

Left-branching language A language where the head of a phrase typically comes after the words that modify its meaning.

Lexical Related to the words of a language and their meaning.

Logographic language A language that uses a writing system where symbols map to the smallest units of sound with meaning called morphemes.

Non-terminal symbol A symbol used in the production rules of a PCFG that maps to a terminal symbol or one or more other non-terminal symbols. An example is the symbol NP which denotes a noun phrase.

Right-branching language A language where the head of a phrase typically comes before the words that modify its meaning.

Surprisal A measure of the unexpectedness of a word given its context, defined as $s_i = -\log_e(P(w_i|w_1, w_2, ...w_{i-1}))$.

Syllabic language A language that uses a writing system where symbols map to the sounds of syllables.

Syntactic Related to the positioning of words to form a sentence in a language.

Terminal symbol A symbol used in the production rules of a PCFG that cannot be mapped to any other symbols. An example is the symbol *cat* which denotes the English noun cat.

Chapter 1

Introduction

Reading is often taken for granted, but the process of constructing meaning from written text is complex and remarkable. Just as astonishing is the vast array of languages and the variety of their writing systems and syntactic structures. These differences influence the mechanisms by which readers understand written texts. Yet many aspects of reading are fundamentally the same across languages (Li et al., 2014). One such example is the anticipation of upcoming words during reading. This capacity for prediction plays a key role in first-language (L1) reading (Huettig, 2015; Kuperberg & Jaeger, 2016; F. Smith, 1975). As observed across a variety of languages, highly predictable words are read more quickly (Ehrlich & Rayner, 1981; Fernández et al., 2014; Kliegl et al., 2004; Rayner & Well, 1996; Yun et al., 2017). However, less is understood about the extent to which prediction plays a role during reading in a second or additional language (L2) (Berzak & Levy, 2022; Chun, 2020; Grüter et al., 2014). Furthermore, the impact of language background when reading in an L2 has not yet been adequately explored. Apart from gaining a deeper understanding of how we process written language, understanding differences in L1 and L2 reading is also critical for informing instructional strategies to support language acquisition.

The predictability of a word can be quantified using surprisal, a measure of the unexpectedness of a word given its context. Surprisal can be computed by language models, of which there are many. Each model has a different way of computing surprisal based on the amount of preceding context used and the degree of syntactic and lexical information included in the training data. In agreement with research on word predictability and reading, surprisal has been found to predict reading times, with low surprisal words (i.e., more predictable words) being read more quickly (Goodkind & Bicknell, 2018; Hale, 2001; Hale et al., 2018; N. J. Smith & Levy, 2013). Although some research has been done regarding the utility of surprisal for predicting L2 reading times (de Varda & Marelli, 2022), there still lacks exploration of language background. One avenue to explore the relationship between language background and L2 reading is to understand how surprisal derived from different language models differentially predicts reading times for people with different language backgrounds. As such, we set out to answer the question: What language models produce surprisals that best predict reading times of English speakers from various language backgrounds?

In the coming chapters, we will provide background about how several language models compute surprisal values and the differences in writing systems and syntactic structures of languages. Next, we will provide more context to explain where our work is situated on the horizons of understanding surprisal's predictive capability in L2 reading. We will describe the data collection procedure of the self-paced reading time data we used, the details of our language model training, and the development of generalized additive mixed models (GAMMs) that use surprisal to predict reading times. We will share our results and discuss our findings in the context of existing knowledge about differences and similarities between L1 and L2 reading as well as language background. Finally, we will conclude by sharing the key insights that answer our research question.

Chapter 2

Background

Language models assign probabilities to sequences of words. These sequences can include phrases, sentences, paragraphs, essays, and longer texts. Language models are useful in natural language processing tasks such as machine translation, speech recognition, question answering systems, and spam detection. They have also been used in psycholinguistics to help us understand sentence processing during reading.

To this end, language models can be used to compute surprisal, a measure of how unexpected a word is given its context. The surprisal s_i of a word w_i in a sentence is defined as $s_i = -\log_e(P(w_i|w_1, w_2, ...w_{i-1}))$. Although this definition remains the same, the surprisals produced from different language models can vary because of the way they estimate the probabilities. The breadth of language models is vast. Here we will describe only those that are relevant to this thesis.

2.1 N-Grams

N-grams are a type of language model that assigns a probability to a word based on the $n-1$ preceding words (Shannon, 1948). In their simplest form, probabilities are calculated based on the occurrence of sequences of words in a training corpus. Several smoothing techniques can be used to account for sequences that do not occur in the training corpus and would otherwise be assigned a probability of zero.

A strength of n-gram language models is interpretability; their simplicity makes them easy to understand. A weakness is their loss of left context; words early in a sequence bear no weight on the probabilities of words that occur late in the sequence (i.e., more than $n - 1$ words later). In other words, n-grams make use of the Markov assumption which states that probabilities can be estimated using only recent information and without making use of information from further in the past.

N-grams compute surprisals based on sequential representations of sentences and do not have access to explicit phrase structure. As such, surprisal values derived from n-grams are primarily informed by lexical information, along with implicit syntactic information. As a variation, syntactic information can be isolated by replacing words with their part-of-speech (POS) tags when training the n-gram model.

2.2 Probabilistic Context-Free Grammars

Probabilistic context-free grammars (PCFGs) were conceptualized in search of a way to appropriately assign probabilities to the words of a language (Booth & Thompson, 1973). They work by assigning probabilities to the production rules of a context-free grammar. Production rules define how the sentences of a language can be constructed. For example, to capture the noun phrase "the cat", a PCFG of English would include a production rule $NP \rightarrow DT\ NN$ and further rules $DT \rightarrow$ *the* and $NN \rightarrow$ *cat*. In this example, *NP*, *DT*, and *NN* are non-terminal symbols while *the* and *cat* are terminal symbols. Non-terminal symbols can be broken down into other non-terminals or a terminal symbol, whereas terminal symbols are final. Of course, there are many other ways to construct a noun phrase in English so the non-terminal *NP* would be involved in several different productions rules. Likewise, *DT* could map to other determiners like "a" and *NN* could map to any singular noun. The complete set of rules of a PCFG can be used to derive a phrase structure tree for a given sentence. From there, the probabilities of the production rules used to construct the tree can be multiplied together to compute the overall probability of the sentence. The probabilities of the production rules are learned by training the PCFG on a large dataset of annotated phrase structure trees.

Several techniques have been explored for using PCFGs to parse sentences and compute surprisal values. Roark et al. (2009) devised a method to use an incremental top-down parser (Roark, 2001) to derive separate syntactic and lexical surprisal values from a PCFG. The syntactic surprisal is calculated based upon the history of derivation including all steps except the final step that maps a POS tag to a word. In this way, the surprisal value does not reflect any lexical information about the upcoming word. The lexical surprisal is defined in such a way that lexical and syntactic surprisal sum to the total surprisal. The total surprisal is calculated based on the entire derivation history, which contrasts the limited context used by n-grams. Beam search is used during parsing to try and obtain the most probable parse and avoid deriving improbable structures.

2.3 Recurrent Neural Network Grammars

Recurrent neural network grammars (RNNGs) were introduced by Dyer et al. (2016) to rectify the incongruity between the purely sequential nature of recurrent neural networks (RNNs) and the inherent hierarchical phrase structure of language. RNNGs can be used to parse an existing sentence into a tree structure or to generate a new sentence along with its tree structure. An RNNG consists of a set of non-terminal symbols, a set of terminal symbols, and a group of neural network parameters that implicitly define the rules of the grammar. These parameters are used by an RNN to select actions based on the current state of the input buffer and stack, which contain the sentence yet to be parsed and the current tree structure, respectively. At each step during parsing, one of the following actions is selected: add a new non-terminal symbol to the stack (NT), shift the next terminal symbol from the input buffer to the stack (SHIFT), or close the newest open non-terminal symbol on the stack to form a completed constituent (REDUCE). The RNN continues to select actions until the input buffer is empty and there is a single completed constituent on the stack. Table

Table 2.1: RNNG parsing example for the input “Forests are disappearing .”

<table>
<tr><th>Stack</th><th>Buffer</th><th>Action</th></tr>
<tr><td></td><td>Forests | are | disappearing | .</td><td>NT(S)</td></tr>
<tr><td>(S</td><td>Forests | are | disappearing | .</td><td>NT(NP)</td></tr>
<tr><td>(S | (NP</td><td>Forests | are | disappearing | .</td><td>NT(NNS)</td></tr>
<tr><td>(S | (NP | (NNS</td><td>Forests | are | disappearing | .</td><td>SHIFT</td></tr>
<tr><td>(S | (NP | (NNS | Forests</td><td>are | disappearing | .</td><td>REDUCE</td></tr>
<tr><td>(S | (NP | (NNS Forests)</td><td>are | disappearing | .</td><td>REDUCE</td></tr>
<tr><td>(S | (NP (NNS Forests))</td><td>are | disappearing | .</td><td>NT(VP)</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP</td><td>are | disappearing | .</td><td>NT(VBP)</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP | (VBP</td><td>are | disappearing | .</td><td>SHIFT</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP | (VBP | are</td><td>disappearing | .</td><td>REDUCE</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP | (VBP are)</td><td>disappearing | .</td><td>NT(VBG)</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP | (VBP are) | (VBG</td><td>disappearing | .</td><td>SHIFT</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP | (VBP are) | (VBG | disappearing</td><td>.</td><td>REDUCE</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP | (VBP are) | (VBG disappearing)</td><td>.</td><td>REDUCE</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP (VBP are) (VBG disappearing))</td><td>.</td><td>SHIFT</td></tr>
<tr><td>(S | (NP (NNS Forests)) | (VP (VBP are) (VBG disappearing)) | .</td><td></td><td>REDUCE</td></tr>
<tr><td>(S (NP (NNS Forests)) (VP (VBP are) (VBG disappearing)) .)</td><td></td><td></td></tr>
</table>

The pipe (|) symbol is used to separate elements of the stack and buffer.

2.1 shows a parsing example. As mentioned, the current state of the stack influences which action is chosen at any given parsing step. This means that actions are chosen based on the entire derivation history of the preceding context. RNNGs are trained on large datasets of phrase structure trees.

RNNGs parse sentences incrementally, which means they are prone to choose actions that seem sensible at the time but are later discovered to be nonsensical upon encountering future words. This risk can be mitigated by using beam search to explore several promising parse options further down the line. However, standard beam search does not work well for RNNGs because structural parse actions are preferred over lexical ones since they are assigned higher probabilities. As such, the RNNG succumbs to complicating the phrase structure of the parse and rarely progresses onward to the next word. Word-synchronous beam search (Stern et al., 2017) addresses this problem by searching through structural actions until a lexical action is eventually chosen. From here, all possible parses have at least reached the next word before beam search proceeds further. This set of parses is pruned based on a given threshold and the resulting set is referred to as the word beam.

Surprisal is calculated over the word beam during the beam search process (Hale et al., 2018). It is computed as the log-ratio of summed forward probabilities. These surprisal values are based on the entire derivation history. As such, they reflect both syntactic and lexical information.

2.4 Transformers

The transformer architecture was first introduced by Vaswani et al. (2017) as a way to overcome the limitations of recurrence and convolution for sequence transformation tasks. They replaced the recurrent and convolutional layers of the best-performing neural network architectures of the time with self-attention mechanisms (Luong et al., 2015) and feed-forward layers. This allows for further parallelization of computation, typically reduces the computational complexity of each layer, and makes it easier to learn long-range dependencies in the sequence. These achievements are due in part to the fact that self-attention layers have direct access to all preceding inputs, whereas recurrent neural networks are limited to the information contained in the current hidden state.

When a transformer is used for computing surprisals, a decoder mask is used to ensure that it does not have access to future items in the sequence. Effectively, this means that the transformer has direct access to all preceding words and the current word but no future words. For example, a transformer would directly use the first, second, and third words of a sentence to compute the surprisal of the fourth word.

Transformers do not explicitly model the hierarchical nature of language since they are trained on sequences of words rather than phrase structure trees. However, attention visualization shows that different attention heads learn to perform different tasks, some related to syntax and some to semantics (Vaswani et al., 2017). As such, the surprisal values computed by a transformer are based on a combination of syntactic and lexical information.

2.5 Language Model Comparison

Table 2.2 outlines the characteristics of each model in a way that is easy to compare and contrast. Of the language models used, the n-gram is the only one that does not use the entire preceding context of a sentence to compute surprisal values. Transformers have the capability to use context spanning multiple sentences, but we train the models such that they only have access to one sentence at a time. The language models also differ by the type of training data they use. N-grams and transformers use sequences of tokens whereas PCFGs and RNNGs use tree structures. Because of this, n-grams and transformers are limited to implicit syntactic information whereas PCFGs and RNNGs have direct access to explicit syntactic information. Because these language models use different amounts of preceding context and are trained on different representations of sentences, the surprisals they produce are informed by different linguistic information.

Table 2.2: Comparison of language models.

	N-Gram	**PCFG**	**RNNG**	**Transformer**
Sentence Context	$n-1$ preceding words	All preceding words	All preceding words	All preceding words
Training Data	Sequences	Tree structures	Tree structures	Sequences
Computational Intensity	Low	Low	High	High
Relative Interpretability	High	High	Low	Low

Apart from these differences in linguistic information, language models differ in resource intensity and interpretability. The computational intensity of training an n-gram or a PCFG is much lower than that required to train an RNNG or transformer. Additionally, understanding exactly how surprisal values are computed is simpler for n-grams and PCFGs whereas it is not as straightforward for RNNGs and transformers.

Chapter 3

Related Work

3.1 Differences in Reading across Languages

Reading is a complex process that involves the incorporation of linguistic knowledge and visually perceived symbols to comprehend text (Woolley, 2011). Many reading mechanisms are universal across languages but others differ based on writing system (Li et al., 2022). Languages are broadly classified as alphabetic, syllabic, or logographic depending on how written symbols correspond to sounds (Anatole, 1997). In alphabetic languages, such as English, symbols often map to the smallest units of sound called phones. Typically each phone is represented by a letter or two. In contrast, syllabic writing systems, use a single symbol to represent a complete syllable. Different yet are logographic languages, such as Chinese, where symbols correspond to morphemes, the smallest units of sound that contain meaning. Beyond the symbolic representation of sounds, writing systems can vary in other ways, such as the direction in which texts are read. Furthermore, some languages mark word boundaries with a space, while others write words contiguously with no explicit demarcation.

Because of these differences, reading in different writing systems is not entirely the same. In alphabetic and syllabic writing systems, the meaning of a word is accessed either directly from orthography or by first accessing the phonology before mapping to meaning (Coltheart et al., 2001; Seidenberg & McClelland, 1989). In contrast, logographic reading appears to rely primarily on the direct link between orthography and meaning, with minimal involvement of phonology (Law et al., 2006; Tan & Perfetti, 1997). There are also processing differences in languages with explicit word boundaries and those without. Readers of spaced scripts tend to fixate on the center of words during reading, but readers of unspaced scripts, such as Chinese, do not exhibit this preference (Li et al., 2011; Rayner, 1979). Furthermore, word boundary ambiguity is sometimes encountered when reading in unspaced scripts and this requires word segmentation, which is not needed for spaced scripts (Li & Pollatsek, 2020).

Beyond writing systems, languages also differ in syntactic structure. For example, the typical sequence that the subject, verb, and object appear in a sentence varies amongst languages and is known as word order. Approximately 80% of languages are subject-verb-object (SVO) or subject-object-verb (SOV) (Dryer, 2013; Hammarström, 2016). English is an SVO language and other examples are Spanish and Chinese. Examples of SOV languages include Korean, Japanese, and Turkish. In addition to word order, languages vary based on branching patterns. In right-branching languages, the head of the sentence usually comes first and is followed by modifiers. Conversely, in left-branching languages, modifiers typically precede the head of the sentence. SVO languages tend to be right branching (e.g., English) while SOV languages tend to left branching (e.g., Korean). Left-branching structures are often more ambiguous since modifiers come first and the meaning may not become clear until the head is encountered. Therefore readers of left-branching languages may need to delay parsing decisions whereas readers of right-branching languages take less risk when making parsing decisions early (Amici et al., 2019). In this way, the word order and branching pattern of a language also play a role in the reading process.

Our discussion so far has focused on how first language reading differs amongst languages, but how do these differences translate to reading in a second language? Research suggests that that previous linguistic knowledge from a person's L1 can transfer to their L2 (Gass, 1979). Further support for this phenomenon can be found in studies showing that L1 and L2 reading skills are positively correlated (Cummins, 1979; Jeon & Yamashita, 2014). As mentioned, individuals develop reading strategies that are specific to the characteristics of their L1 (McNeill et al., 1971). It seems that these strategies are applied to an extent when reading in their L2 (Koda, 1990). This implies that L1 and L2 speakers may employ different reading strategies for the same language depending on their linguistic background.

3.2 Similarities in Reading across Languages

Despite differences between writing systems, there are shared fundamentals of reading across languages (Li et al., 2014). Evidence suggests that words are the basic unit of processing even in writing systems that do not explicitly mark word boundaries. Reading appears to be incremental, with readers immediately incorporating each word as it is encountered to construct the complete meaning of a sentence. It has been observed that higher frequency and shorter words are read more quickly and that prediction plays an important role during reading (Huettig, 2015; Kuperberg & Jaeger, 2016; F. Smith, 1975). Experiments by Ehrlich and Rayner (1981) showed that highly predictable words are less likely to be fixated on during reading and less time is spent reading them when they are fixated. These results have been corroborated by other studies that also observed reduced fixation time for predictable words (Rayner & Well, 1996). Although much of the existing research that investigates the influence of word predictability on reading focuses on English, there have been several studies that observe a similar phenomenon in other languages, e.g., Arabic (AlJassmi et al., 2022); Chinese (Rayner et al., 2005); German (Kliegl et al., 2004); Korean (Yun et al., 2017); and Spanish (Fernández et al., 2014).

3.3 Language Models for Predicting Reading Times

Aligned with research on word predictability and reading, there is an abundance of research showing that surprisal is a strong predictor of reading time for English (Goodkind & Bicknell, 2018; Hale, 2001; Hale et al., 2018; N. J. Smith & Levy, 2013). Surprisal is a measure of how unexpected a word is given its context and can be computed by language models. More formally, surprisal is the negative logarithm of the probability of a word's occurrence conditioned on the preceding words. A low surprisal value corresponds with high predictability. Thus, words with low surprisal values are generally read more quickly whereas words with high surprisal are read more slowly. There are many language models and each has a different approach for computing surprisal. Refer back to Chapter 2 for a discussion about how surprisals are calculated by n-gram, PCFG, RNNG, and transformer language models.

Although the relationship between surprisal and reading time is well established for L1 reading, the role of prediction in L2 reading is less conclusive. Grüter et al. (2014) found that even the most proficient L2 readers have a reduced capacity to predict upcoming words during sentence processing compared to their L1 counterparts. They reasoned that L2 processing is exhausted by reactive processing, leaving little room for prediction. However, other researchers (Berzak & Levy, 2022) found that the degree to which L2 processing relies on prediction is dependent on proficiency, with the most proficient L2 readers using similar strategies to L1 speakers. Other research corroborates that L2 readers incorporate prediction into processing and also suggests that previous linguistic knowledge impacts the predictions these readers make (Chun, 2020).

Despite interest in the role of prediction in L2 reading, limited research has been conducted on whether surprisal predicts L2 reading times. A study by de Varda and Marelli (2022) explored this open question using the recently released Multilingual Eye-Movements Corpus (Kuperman et al., 2023; Siegelman et al., 2022). Using a linear mixed effects model, they found that an effect of surprisal on L2 reading times is present but to a lesser extent than for L1 reading. However, several questions regarding the relationship between surprisal and L2 reading times remain unanswered. In particular, research lacks an investigation of the relationship between language background and the predictive capability of surprisal. Additionally, the wide variety of language models available for computing surprisal have not been adequately explored for predicting reading times. Given these research avenues, we set forth to determine what language models produce surprisals that best predict reading times for people with different language backgrounds. In this pursuit, we hope to better understand the relationship between language background and reading by drawing insights from the ways in which the best performing surprisals were calculated.

Chapter 4

Methods

We trained nine different language models that varied in the amount of syntactic information, lexical information, and preceding context they used to compute surprisal. We used these surprisal values to predict the English reading times of people with different language backgrounds. For each language background, we developed a GAMM containing covariates other than surprisal that are known to predict reading time. We then developed nine more GAMMs for each participant group containing these covariates plus terms for surprisal derived from one of the language models. We compared the models with surprisal to the one without on the basis of Akaike information criterion (AIC) to determine which surprisal provided the most useful information for predicting English reading times of people with different L1s.

4.1 Reading Time Data

The reading time data used in this work were collected by collaborators from the University of Pittsburgh. Max Helfrich collected the data for a project led by Lin Chen and Charles Perfetti. Participants from various language backgrounds were asked to read a series of passages in a self-paced moving window paradigm where reading time was recorded as the time between key presses. The following sections detail the participants, stimuli, and data collection procedure of the study.

4.1.1 Participants

Participants included adults with different first languages (English, $n = 34$; Chinese, $n = 35$; Korean, $n = 28$; Spanish, $n = 44$) who speak English. It is possible that the participants speak additional languages other than English and their first language; this information was not collected.

Table 4.1: Gender and age of participants by first language.

		English	**Chinese**	**Korean**	**Spanish**
Gender (M:F:O)		3:30:1	19:16:0	11:17:0	15:32:2
Age	Range	17-28	18-22	18-25	18-32
	Average	18.6 (*SD* = 2.0)	19.4 (*SD* = 1.4)	19.6 (*SD* = 2.2)	19.2 (*SD* = 2.5)

Options for gender included male (M), female (F), and other (O).

4.1.2 Participants' First Languages

Our research focuses on evaluating the performance of surprisals produced by various language models for the predicting reading times of people with different language backgrounds. To be able to glean meaningful insights from the interplay between the characteristics of the reader's first language and how the best performing surprisals were computed, it would be desirable to consider a wide array of languages with varied linguistic features. Participants of this study were native speakers of English, Chinese, Korean, or Spanish. Although only four language backgrounds were explored, these four languages have features distinct from one another. Table 4.2 outlines the differences in language family, writing system, word boundary marking, word order, and branching tendency amongst these languages. Some of these differences will be relevant to the discussion of our results later on.

Table 4.2: Comparison of English, Chinese, Korean, and Spanish.

	English	**Chinese**	**Korean**	**Spanish**
Language Family	Germanic	Sino-Tibetan	Altaic	Romance
Writing System	Alphabet	Logography	Alphabetic Syllabary	Alphabet
Word Boundaries	Marked	Unmarked	Marked	Marked
Word Order	SVO	SVO	SOV	SVO
Branching	Right	Left	Left	Right

4.1.3 Measures of Participants' English Language Skills

The participants' English language skills were measured by the Nelson-Denny vocabulary and reading comprehension tests (Brown et al., 1993). The vocabulary test had 100 multiple-choice items and the reading comprehension test had 36 multiple-choice items. The L1 English speakers were given 7.5 minutes to complete the vocabulary test and 15 minutes to complete the reading comprehension test. The L2 English speakers were given extended time with 15 minutes for the vocabulary test and 30 minutes for the reading comprehension test.

The test scores were reported as performance and adjusted performance. Performance was the score out of all questions on the test whereas adjusted performance was the score out of only the questions that were attempted.

4.1.4 Stimuli

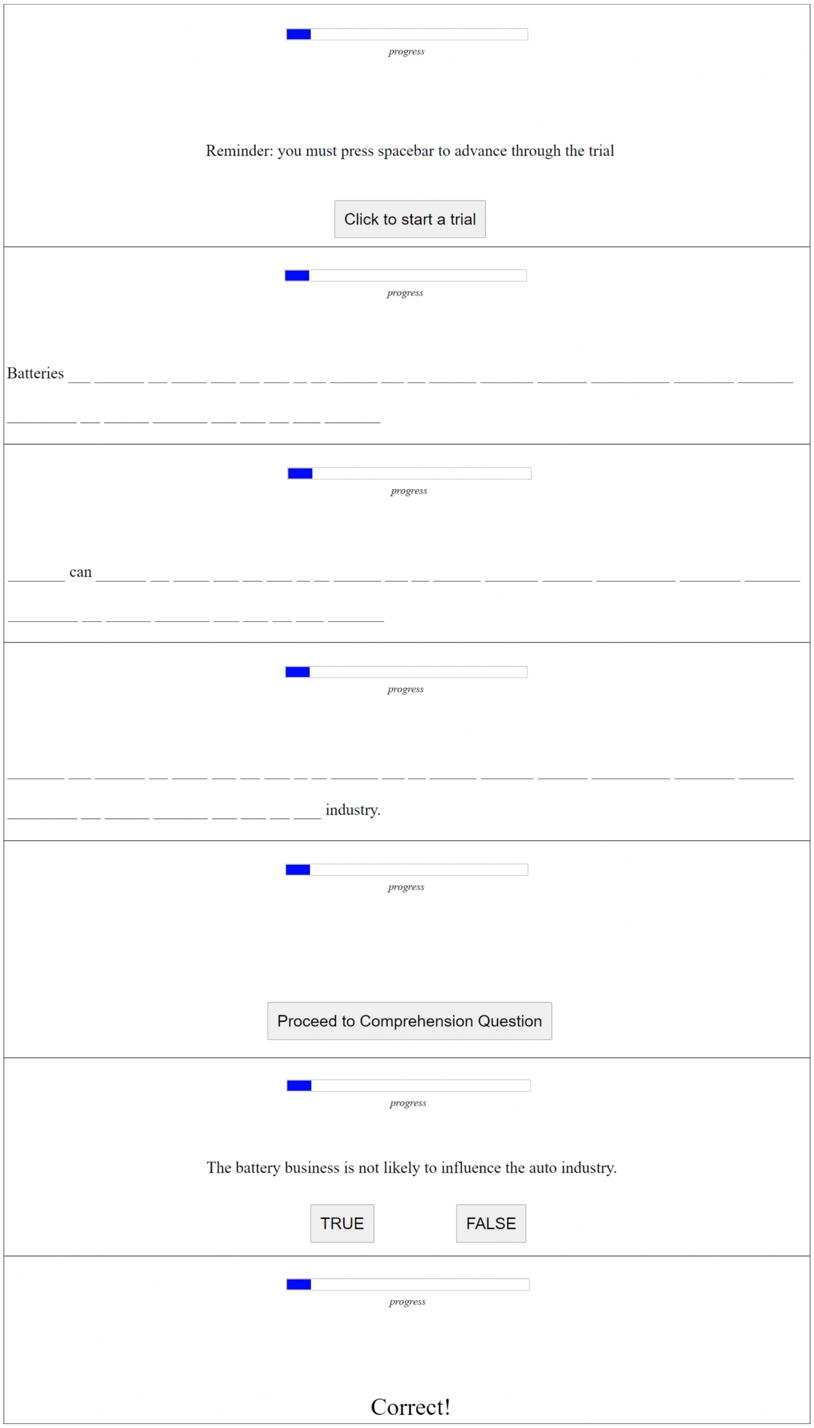
The stimuli consisted of 63 two-sentence passages from articles of the New York Times published in 2018 and 2019. As a popular newspaper, the New York Times is a source for texts with style and content likely familiar to the participants. When selecting the passages, politically and culturally sensitive topics were avoided. As a whole, the chosen topics were intended to appeal to a wide variety of people and relate to general knowledge, avoiding niches and jargon. The passage length was limited to two sentences to prevent participant fatigue.

The passages were 11 to 30 words long with an average length of 22.4 (*SD* = 5.2). The sentences making up these passages ranged in length from 3 to 22 words with an average length of 11.2 (*SD* = 4.5). An example of a passage is "Forests are disappearing. Maps show shrinking woodlands all over the world." Appendix A provides a full list of the stimuli.

4.1.5 Data Collection

The study was conducted using PCIbex (Zehr & Schwarz, 2018) where the stimuli were presented in a self-paced moving window paradigm (see Figure 4.1). Each passage appeared as a series of black dashes on a white background that indicated the length and position of each word in the passage. When the participants were ready, they pressed the space bar to reveal the first word. Once they had read that word, they pressed the space bar again to reveal the next word and revert the first one back to a dash. The reading time for each word was recorded as the time between these keystrokes. This process was repeated until the entire passage had been read.

At this point, participants were presented with a true or false reading comprehension question (if the passage had one) to check their attentiveness. Two-thirds of the passages had an associated question and half the answers were true. Participants answered these questions by clicking a button and were given feedback on screen that let them know whether their answer was correct. After each passage, participants had the option to take a break or directly move on to the next passage. Participants received reminders if they did not respond within 30 s during the passage reading phase or within 10 s during the question answering phase. The order of the passages was random for each participant. Commas and periods were presented with the preceding word. Before data collection began, each participant had six practice passages and four of these had a reading comprehension question.



Reminder: you must press spacebar to advance through the trial

Batteries

can

industry.

The battery business is not likely to influence the auto industry.

Correct!

Figure 4.1: Self-paced moving window setup.

4.1.6 Data Cleaning

We prepared the reading time data for statistical modelling. We removed data for which the reading comprehension question was answered incorrectly (English, 4.6%;

Chinese, 10.9%; Korean, 9.5%; Spanish, 8.0%). Following standard practice in the field (Tremblay et al., 2011), reading times outside of ±2.5 *SD* for each participant were removed (English, 2.1%; Chinese, 1.9%; Korean, 1.9%; Spanish, 1.9%). Following Monsalve et al. (2012), reading times less than 50 ms or greater than 2,000 ms were removed (English, 3.1%; Chinese, 0.5%; Korean, 0.4%; Spanish, 0.3%).

We performed a log transform on the reading times to make the data more normally distributed. After this, we checked the distributions by participant and removed any that had unusual patterns that suggested there was some sort of error during data collection. We removed one participant's data based on this check. This participant was a Korean native speaker; their data was also excluded from the demographic information reported in Table 4.1.

4.2 WikiText-2 Data

4.2.1 Dataset Description

We trained the language models using the WikiText-2 dataset (Merity et al., 2017), which contains articles from Wikipedia. The dataset only includes articles that have been verified as Good or Featured by the editors of Wikipedia. The articles cover a wide variety of topics ranging from geography and history to pop culture and games. The dataset includes a train split with 600 articles along with validation and test splits with 60 articles each. We used the version that did not have any tokens replaced with <unk>. It can be downloaded from https://huggingface.co/datasets/wikitext.

4.2.2 Preprocessing

We preprocessed these splits to prepare the data for training. The articles were split into sentences using NLTK *sent_tokenizer* (Bird et al., 2009). Since the transformer requires a maximum sentence length to be specified, sentences longer than 52 tokens were removed from the data. Before removal, sentences containing semicolons were split into two separate sentences in an effort to preserve as much data as possible.

The maximum length was chosen to be 52 since it balanced the computational cost of training the transformer with the preservation of training data. This value preserved 98% of the data. Additionally, the longest sentence in the stimuli had a length of 22 tokens, so excluding sentences longer than 52 tokens does not limit the training data from being representative of the stimuli.

Most other preprocessing decisions were made on the basis of aligning the WikiText-2 data more closely with the stimuli. First, article headings and subheadings were removed from the data. We excluded these because they were typically phrases, but the stimuli only contained full sentences. Instances of “can not” were replaced with “cannot” to match what participants saw in the stimuli. Along the same lines, contractions were split into full words using the contractions Python library (van Kooten, 2022) since the stimuli do not contain any contractions. All tokens containing punctuation or symbols other than commas, periods, and the apostrophe for the possessive “s” were removed since the stimuli only contain commas and periods.

Five versions of the data were required for the various language models: word, POS, combined word and POS, lexicalized tree, and unlexicalized tree. The POS version was generated by POS tagging the sentences using the fine-grained tag set of spaCy (Honnibal & Montani, 2017). Sentences containing the POS tags :, ‘, and -LRB-, -RRB- were manually inspected as these should not have been present since such punctuation was removed. This led the the removal of one sentence which appeared to be a chemical equation and the correction of the POS tags for another sentence. The word version was created by lowercasing all words to match the stimuli and replacing a word with its POS tag if it was CD (i.e., the POS tag for numerals and cardinals in written and numerical form). We did this to avoid a sparse representation of numbers in the model and because we were not interested in the number itself but rather the fact that it is a number. The combined word and POS version was created by simply conjoining each word with its POS tag separated by a forward slash (e.g., exercise/NN).

The lexicalized tree version was constructed using the Stanford Parser version 4.2.0 (Klein & Manning, 2003) with the tags from the spaCy tagger provided alongside the words as input. The unlexicalized trees were constructed from these trees by simply replacing the words with their POS tags. Table 4.3 shows the descriptive statistics for the data splits of the WikiText-2 dataset after preprocessing.

Table 4.3: Data splits after preprocessing the WikiText-2 dataset.

	Train	**Validation**	**Test**
Articles	600	60	60
Sentences	79,109	8,219	9,716
Tokens	1,812,933	189,795	215,346

4.3 Language Models

4.3.1 N-Gram

We used KenLM (Heafield, 2011) to train n-gram language models on the preprocessed WikiText-2 data. We trained three models using different versions of the dataset: word only, POS only, and combined word and POS. We performed hyperparameter tuning for each of the models on the value of n in the range of 2 to 6, which is the full range that is supported by KenLM. We selected the values of n that minimized the word perplexity on the validation set. This resulted in values of 6, 5, and 6 for the word only, POS only, and combined word and POS versions of the dataset, respectively. We evaluated the models on the test set using word perplexity. Training was performed on an Apple M1 Pro chip. The hyperparameter tuning process required training and evaluating 15 models (5 values of n × 3 dataset versions). This entire process took 41 seconds total of computation time.

4.3.2 PCFG

We used the incremental top-down parser of Roark et al. (2009) to train a PCFG language model on the preprocessed WikiText-2 data. Similarly to Demberg and Keller (2008), we trained two models: one using the lexicalized trees and one using the unlexicalized trees. We performed hyperparameter tuning on the threshold, testing values of 0.02, 0.04, 0.05, 0.06, and 0.08. We selected the threshold that minimized the word perplexity on the validation set. This was found to be 0.02 for both the lexicalized and unlexicalized treees. We evaluated the models on the test set using word perplexity. Training was performed on an Apple M1 Pro chip. The hyperparameter process required training and evaluating ten models (5 values for the threshold × 2 dataset versions). This entire process took one hour total of computation time.

4.3.3 RNNG

We trained an RNNG on the lexicalized tree version of the preprocessed WikiText-2 data using the pyTorch implementation of Noji and Oseki (2021). We used this implementation to exploit GPU parallel computing to reduce the training time. We did not train an RNNG using an unlexicalized version of the data because training is resource intensive.

We performed tuning on several hyperparameters including learning rate, dropout rate, batch size, and batch composition. Due to the number of hyperparameters, the search space was quite expansive and it was not feasible to test all combinations of all hyperparameters. As such, the hyperparameter tuning process was split into two rounds (see Table 4.4). The first round was a coarse-grained search of the four hyperparameters covering a large range of potential values. The hyperparameter values that minimized the perplexity on the validation set for this round informed the refinement of the search space for the second round. The final hyperparameter values for the RNNG were based on the results of the second round. The final values were a learning rate of 0.002, dropout rate of 0.2, batch size of 256, and batch composition of random length sentences. Training was performed using an Intel Silver 4216 Cascade Lake CPU, NVIDIA V100 Volta GPU, and 32 GB of memory. Training for the complete hyperparameter tuning procedure was completed in just under 7 days.

Table 4.4: Values included in the hyperparameter tuning process for the RNNG.

	Learning Rate	**Dropout Rate**	**Batch Size**	**Batch Composition**
Round 1	0.0001	**0.1**	**256**	similar length
	0.001	0.5	512	**random length**
	0.01	0.9	1024	
Round 2	**0.002**	0	128	**random length**
	0.01	0.1	**256**	
	0.05	**0.2**		
		0.3		
		0.4		

All combinations of the listed parameter values were tested for each round. Bold text indicates the parameters that minimized the validation word perplexity for each round.

The configuration of the RNNG after hyperparameter tuning was as follows. The RNNG employed a two-layer LSTM with hidden dimensions of 256. The input and output word dimensions were also 256. A dropout rate of 0.2 was applied. The parameters were initialized to zero before training and the training sentences were randomly grouped into batches of 256. During training for 18 epochs, the parameters were optimized using Adam and cross-entropy loss with a constant learning rate of 0.002. Gradient clipping was used to prevent the problem of exploding gradients and a maximum gradient norm of 5 was used. The random seed was set to 3435 for reproducibility. The total number of model parameters was 18,119,759.

Figure 4.2 shows the training and validation word perplexity over the training epochs. Note that the lines for training and validation perplexity overlap at some points and in these cases they appear as a single grey line. Epoch 0 is before training. The validation word perplexity plateaus by the ninth epoch, indicating that the number of epochs was adequate. Additionally, the model does not appear to be overfitting to the training data since the validation perplexity does not increase with further training after the plateau.

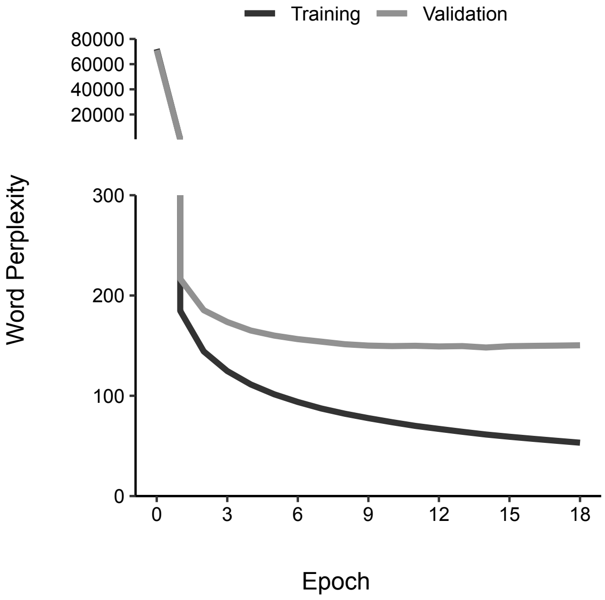


Figure 4.2: Training and validation word perplexity over training epochs for the RNNG with the selected hyperparameters.

The settings for beam search were also tuned. In particular, block sizes of 50, 100, and 200 were tested in combination with beam sizes of 100, 200, 400, 600, and 800. Batch size was set to be the block size divided by 10, word beam size was set to be the beam size divided by 10, and shift size was set to be the beam size divided by 100, following previous research (Hale et al., 2018; Noji & Oseki, 2021; Stern et al., 2017). Validation perplexity plateaued at a beam size of 400 while block size had a very minimal effect on performance. As such, the surprisal for the stimuli were computed using a block size of 100, batch size of 10, beam size of 400, word beam size of 40, and shift size of 4. Beam search was performed using an Intel Silver 4216 Cascade Lake CPU, NVIDIA V100 Volta GPU, and 64 GB of memory. It took 2.5 days to complete the hyperparameter tuning and an additional maximum of 2 days to obtain the surprisal values for the stimuli and test set.[1]

[1] This was the maximum time allowed but the actual time to complete the job was likely less. An exact duration cannot be provided since the job ID was lost.

4.3.4 Transformer

We trained a transformer following Merkx and Frank (2021). This model used the word only version of the preprocessed WikiText-2 data. We did not train the model using the POS version nor the combined word and POS version due to resource intensity. Merkx and Frank (2021) tested two configurations for the transformer and found that the two-layer architecture was better at explaining human reading data. As such, we trained a two-layer transformer. We performed hyperparameter tuning on the learning rate and batch size (see Table 4.5). A learning rate of 0.005 and a batch size of 5 were selected. Hyperparameter tuning was performed using an Intel Silver 4216 Cascade Lake CPU, NVIDIA V100 Volta GPU, and 32 GB of memory. Training for the complete tuning procedure was completed in approximately 7 hours.[2] Evaluating the models on the validation set was performed using an Intel Silver 4216 Cascade Lake CPU, NVIDIA V100 Volta GPU, and 64 GB of memory. The validation set was split into six equally sized subsections for evaluation with each portion evaluated within 3 days.[3]

Table 4.5: Values included in the hyperparameter tuning process for the transformer.

Learning Rate	**Batch Size**
0.0002	**5**
0.001	10
0.005	20
0.025	40
0.125	80

All combinations of the listed parameter values were tested. Bold text indicates the values that minimized the validation word perplexity.

The configuration of the transformer after hyperparameter tuning was as follows. The model had an embedding layer with 400 dimensions per word, followed by two

[2] Training with the selected hyperparameters took 16.5 minutes and 25 models were trained.

[3] This was the maximum time allowed but the actual time to complete the job was likely less. An exact duration cannot be provided since the job IDs were lost.

transformer layers with 8 attention heads and a fully connected layer with 1024 units each. The maximum length was set to 54, which accommodates a sentence with 52 tokens, allowing two tokens for sentence start and end markers. A dropout rate of 0.1 was used. The transformer parameters were randomly initialized with the random seed set to 745546129 for reproducibility. Training sentences were randomly grouped into batches of size 5. Stochastic gradient descent was used to optimize the parameters over 8 training epochs based on cross-entropy loss and using a learning rate of 0.005. The total number of model parameters was 11,952,006.

Figure 4.3 shows the training and validation word perplexity over the training epochs. Note that the lines for training and validation perplexity overlap at some points and in these cases they appear as a single grey line. Epoch 0 is before training. The number of training epochs seems suitable since the validation perplexity plateaus after four epochs and does not increase afterward.

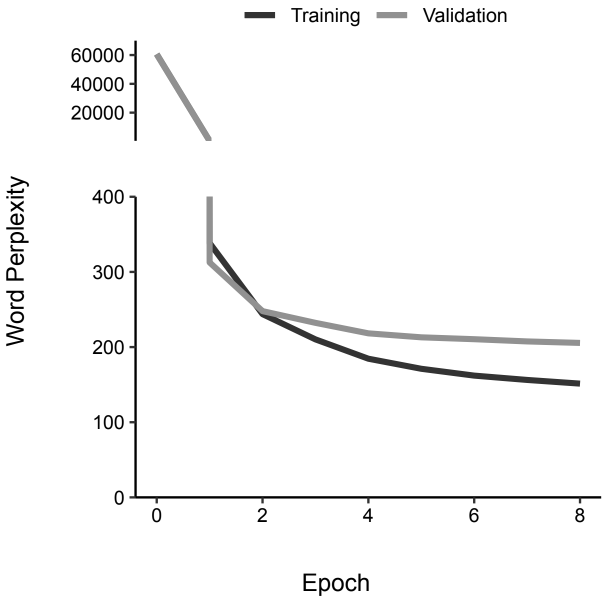


Figure 4.3: Training and validation word perplexity over training epochs for the transformer with the selected hyperparameters.

4.4 Evaluation

All language models were evaluated on the test split of the preprocessed WikiText-2 dataset in terms of perplexity. Perplexity is defined as $ppl = e^{\frac{1}{n}\sum_{i=1}^{n} s(w_i)}$, where n is the number of words and $s(w_i)$ is the surprisal of word i. Note that the direct mathematical relationship between surprisal and perplexity indicates that perplexity could be used in place of surprisal for predicting reading times, at least when using statistical models that can capture non-linear relationships between variables. Despite this observation, and the fact that language models and perplexity are highly associated, we chose to use surprisal as the predictor of reading time for consistency with previous work.

4.5 Generalized Additive Mixed Models

We used GAMMs to predict reading time from a variety of covariates and surprisal. GAMMs allow us to model complex non-linear relationships between the independent variables and the dependent variable. They balance the predictive power of black-box machine learning models with the interpretability of statistical models (S. Wood, 2006).

4.5.1 Base Models

For each language background, we trained a base model that included factors other than surprisal that are known to affect reading time. In particular, the covariates included in the base model were:

- A random effect for participant.
- A random effect for trial. Trial refers to how many passages the participant has read (e.g., the first passage that a participant read is trial 1).

- A random effect for word. Word refers to the lowercase version of the word as a factor.
- A smooth fixed effect for adjusted performance on the vocabulary test.
- A smooth fixed effect for adjusted performance on the reading comprehension test.
- A linear fixed effect for sentence position. Sentence position indicates whether the word was in the first or second sentence of a passage.
- A smooth fixed effect for word position. Word position indicates the position of the word in a sentence (e.g. the first word of a sentence has a word position of 1).
- A smooth fixed effect for log word frequency. Log word frequency was computed from the preprocessed WikiText-2 training data using a base 10 logarithm.
- A smooth fixed effect for word length. Word length refers to the number of letters in a word.
- A tensor product interaction term between log word frequency and word length.

The random effect for participant was included to take into account the fact that each participant has a different mean reading time for reasons that cannot be easily measured. Including trial accounted for effects on reading time stemming from fatigue or increasing familiarity with the task. The random effect for word accounted for differences in the time it takes to read a particular word that cannot be explained by the other features of a word like word length and word frequency.

Vocabulary and reading comprehension adjusted performance scores were included in the model to account for the effect of English language proficiency on reading time. Sentence position was included to account for differences across sentences. This is important considering that the word surprisals were calculated using only the current sentence as context. However, the participants read two sentences consecutively, so the content of the first sentence would influence their expectations for the second sentence. In this way, there was a disconnect between the computed surprisals and human-like expectation. Including an effect for sentence position was intended to account for this incongruity. Word position was included because previous research has shown that it impacts reading time (Kuperman et al., 2010). Effects for log word frequency, word length, and the interaction between them were included following previous research (Goodkind & Bicknell, 2018; N. J. Smith & Levy, 2013). Log word frequency is correlated with surprisal, but differs from surprisal in that it is context independent. Words that are more frequent would be expected to be more familiar to the reader, and therefore take less time to read. Word length affects reading time due to the additional time needed to visually process longer words. Since these features were expected to affect reading time, they must be included in the model to avoid omitted variable bias.

The models were fit using the bam() function in mgcv (S. N. Wood, 2004) with the maximum likelihood method for finding the smoothing parameter. All of the random effects listed above were modeled using a factor smooth basis (i.e., bs="fs" in the bam() function call). This smooth is intended to be used for modelling random effects for factors with many levels, making it appropriate for our data. All of the smooth fixed effects were modeled using thin-plate regression splines, the default basis for smooths in mgcv. The tensor product interaction term was modeled separately from the main effects (i.e., using ti() rather than te()) to allow log word frequency and word length to be used in multiple interactions without the main effects being duplicated. We fit the models using the default number of basis functions for the smooth terms and then used the k.check() function to confirm that it was adequate for capturing the true pattern of the data. The four base models were fit simultaneously using four CPU cores of an Apple M1 Pro chip in approximately 25 minutes.

4.5.2 Full Models

For each participant group, nine more GAMMs were created in addition to the base model: one for each of the surprisal types derived from the different language models. Each full GAMM consisted of the base model plus these additional terms including surprisal:

- A smooth fixed effect for surprisal.
- A linear interaction effect between surprisal and sentence position.
- A tensor product interaction term between surprisal and word position.
- A tensor product interaction term between surprisal and log word frequency.
- A tensor product interaction term between surprisal and word length.

In addition to the smooth term for surprisal, we included several terms for interactions between other features of the word and surprisal. Since surprisals were computed per sentence, the context of the first sentence was not reflected by the surprisal values of words in the second sentence. As such, we included an interaction term between surprisal and sentence position to account for words in the second sentence being more predictable than what was captured by their surprisal values. We similarly included an interaction between surprisal and word position to account for increased predictability towards the end of a sentence that is not captured by language models that use limited preceding context, such as n-grams. Since we included an interaction between log word frequency and word length in the base model, we also included interactions between these factors and surprisal since log word frequency and surprisal are correlated.

We followed the same fitting procedure as for the base model. In particular, we fit these models using the bam() function in mgcv (S. N. Wood, 2004) with the maximum likelihood method for finding the smoothing parameter. The smooth fixed effects were modeled using thin-plate regression splines and the tensor product interactions were modeled separately from the main effects. We used the default number of basis functions for fitting and verified this was enough using k.check(). The full models were fit using six CPU cores of an Apple M1 Pro chip and the entire fitting process was completed in 3.5 hours. During this time, a total of 36 full models (9 surprisal types × 4 language backgrounds) were trained.

4.5.3 Model Comparison

We used AIC to compare the full models with surprisal terms to the base model for each participant group. AIC is an estimate of prediction error so the lower the value, the better. It also takes into account model complexity, with complex models penalized over simpler models. The absolute value of AIC is not informative, but the difference in AIC values between models is. We defined $\Delta AIC = AIC_{\text{base model}} - AIC_{\text{full model}}$ to measure the increase in model performance with the addition of surprisal terms. There is no test to determine if there is a statistically significant difference in ΔAIC values. However, a difference in AIC of 2 or more is typically considered significant following a guideline (Burnham & Anderson, 2003). Although this guideline is empirically based, Burnham and Anderson (2003) provided further support for it by deriving the guidelines using three different approaches. We used ΔAIC to compare the various surprisals to determine which are best at predicting reading times of people with different language backgrounds.

Chapter 5

Results

First we will present the participants' lengths of residence in the U.S. followed by their scores on the vocabulary and reading comprehension tests. Then we will report the performance of each language model on the test set in terms of perplexity. Next we will illustrate the differences in distributions among the surprisal values for the stimuli by language model. Finally, we will report the significance of the terms included in the GAMMs and compare the performance of the various surprisals for predicting reading times of people with different language backgrounds.

5.1 Participant Length of Residence in U.S.

Figure 5.1 shows the participants' lengths of residence in the U.S. by first language. A Kruskal-Wallis test found a significant difference in length of residency for participant's from different language backgrounds ($H = 26.23$, $p < .001$, $\eta^2 = .23$). Table 5.1 shows the results of post-hoc pairwise Mann-Whitney tests with Holm correction. Holm correction was used to limit the probability of false positives due to multiple comparisons. These results indicate significant differences in length of residence in the U.S. between the participants who are L1 speakers of Chinese, Korean, and Spanish.

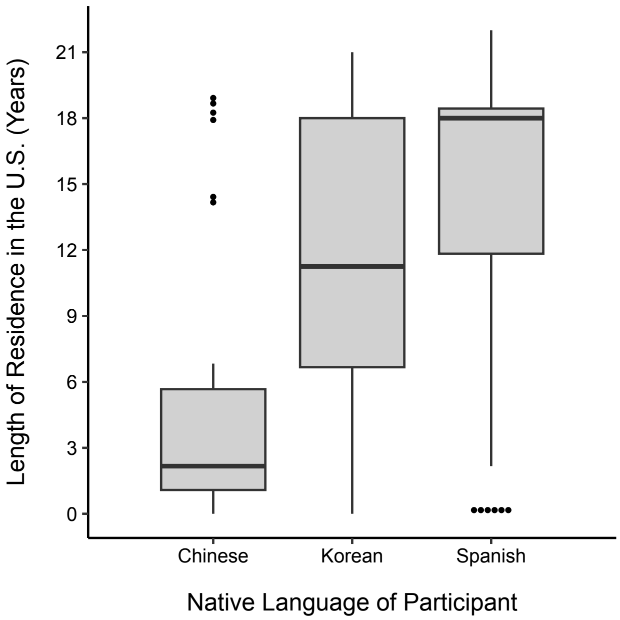


Figure 5.1: Length of residence in the U.S. by first language.

Table 5.1: Two-tailed Mann-Whitney test results comparing length of residence in the U.S. between participants with different first languages.

		n_1	n_2	U	p	r
Chinese	Korean	35	28	204	$< .001$	.50
Chinese	Spanish	35	44	316	$< .001$	.51
Korean	Spanish	28	44	443	.044	.24

5.2 Participant English Proficiency

Figure 5.2 shows the English proficiency scores of the participants as measured by the Nelson-Denny reading comprehension test and vocabulary test (Brown et al., 1993). Performance was the score out of all questions on the test and adjusted performance was the score out of the questions that were attempted.

A Kruskal-Wallis test found a significant difference in English reading comprehension test adjusted performance of participants from different language backgrounds ($H = 10.39$, $p = .016$, $\eta^2 = .05$). Table 5.2 shows the results of post-hoc pairwise Mann-Whitney tests with Holm correction. These results indicate significant differences between the participants who are L1 speakers of English and Chinese. No other differences were found.

Table 5.2: Two-tailed Mann-Whitney test results comparing English reading comprehension test performance between participants with different first languages.

		n_1	n_2	U	p	r
English	Chinese	34	25	856	.010	.38
English	Korean	34	28	530	1.000	.10
English	Spanish	34	44	833	1.000	.10
Chinese	Korean	35	28	350	.210	.25
Chinese	Spanish	35	44	530	.089	.27
Korean	Spanish	28	44	624	1.000	.01

A Kruskal-Wallis test found a significant difference in English vocabulary test adjusted performance of participant's from different language backgrounds ($H = 13.24$, $p = .004$, $\eta^2 = .07$). Table 5.3 shows the results of post-hoc pairwise Mann-Whitney tests with Holm correction. These results indicate a significant difference between the participants who are L1 speakers of English and Chinese as well as L1 speakers of Korean and Chinese. No other differences were found.

Table 5.3: Two-tailed Mann-Whitney test results comparing English vocabulary test performance between participants with different first languages.

		n_1	n_2	U	p	r
English	Chinese	34	25	874	.005	.40
English	Korean	34	28	501	1.000	.04
English	Spanish	34	44	873	.629	.14
Chinese	Korean	35	28	299	.042	.33
Chinese	Spanish	35	44	520	.055	.28
Korean	Spanish	28	44	672	1.000	.08

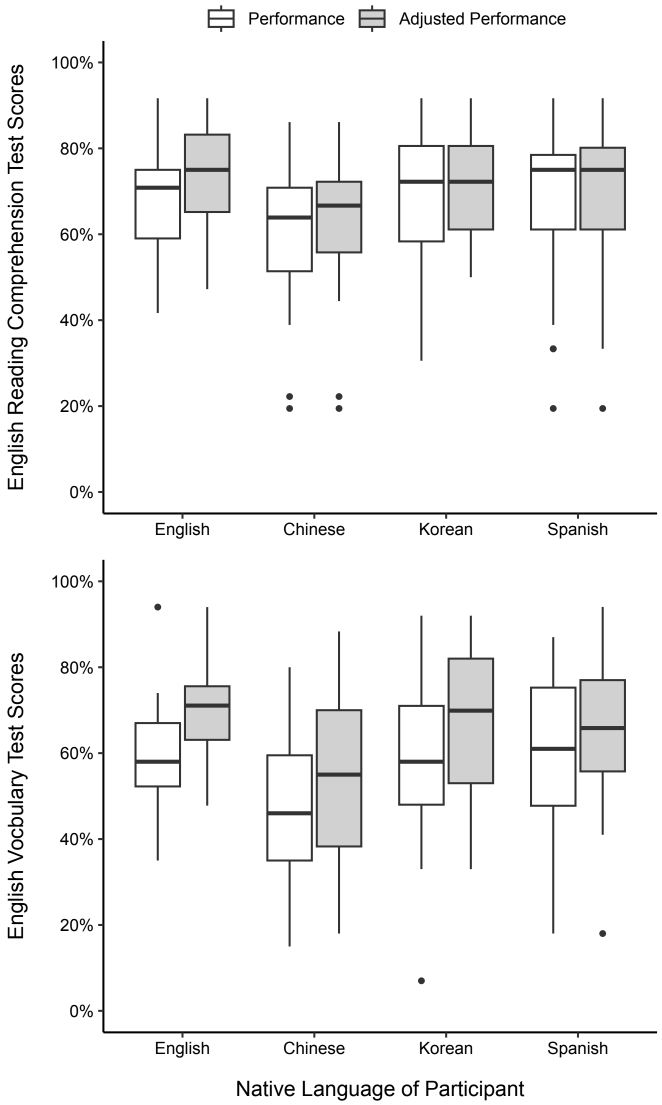


Figure 5.2: English proficiency scores by first language.

5.3 Language Model Perplexities on Test Set

Table 5.4 shows each language model's perplexity on the WikiText-2 preprocessed test set. Direct comparisons should not be made between models that differ by training data type (i.e., word, POS, word/POS, lexicalized trees, and unlexicalized trees). Additionally, the PCFG syntactic and lexical perplexities should not be compared with other models because the surprisals from these models are broken down from the PCFG total surprisal. Between the two sequential models, the transformer better captured the training data as evidenced by its low perplexity in comparison to the n-gram word model. For the hierarchical models, the PCFG had a lower perplexity than the RNNG.

Table 5.4: Perplexities of the various language models on the test set.

Language Model	**Perplexity**
N-Gram Word	398.6
Transformer	310.5
PCFG Total	341.7
RNNG	402.8
N-Gram POS	7.7
N-Gram Word/POS	451.5
PCFG Syntactic	4.5
PCFG Lexical	76.0
PCFG POS	7.6

Language models within the same division were trained on identical versions of the data.

5.4 Stimuli Surprisals

Figure 5.3 shows violin plots of the surprisals derived from each of the language models for the stimuli. The models that use only syntactic information to compute surprisal were plotted separately from those that explicitly include lexical information due to the difference in scale that made it difficult to see the shapes of all plots when plotted together.

A Kruskal–Wallis test found a significant difference in surprisal values computed by various language models ($H = 4936.80$, $p < .001$, $\eta^2 = .39$). Table 5.5 shows the results of post-hoc pairwise Mann-Whitney tests with Holm correction. All groups have a sample size of 1409.

Table 5.5: Two-tailed Mann-Whitney test results comparing stimuli surprisal values computed by different language models.

		U	***p***	***r***
N-Gram Word	N-Gram POS	1737121	$< .001$	.65
N-Gram Word	N-Gram Word/POS	969229	.985	.02
N-Gram Word	PCFG Total	1020506	.985	.02
N-Gram Word	PCFG Lexical	1265676	$< .001$	.24
N-Gram Word	PCFG Syntactic	1834123	$< .001$	.73
N-Gram Word	PCFG POS	1731763	$< .001$	.64
N-Gram Word	RNNG	1062201	.013	.06
N-Gram Word	Transformer	1038688	.200	.04
N-Gram POS	N-Gram Word/POS	255780	$< .001$	.64
N-Gram POS	PCFG Total	192440	$< .001$	.70
N-Gram POS	PCFG Lexical	460501	$< .001$	.46
N-Gram POS	PCFG Syntactic	1289034	$< .001$	.26
N-Gram POS	PCFG POS	975256	.985	.02
N-Gram POS	RNNG	288160	$< .001$	.61
N-Gram POS	Transformer	297433	$< .001$	.61
N-Gram Word/POS	PCFG Total	1042151	.175	.04
N-Gram Word/POS	PCFG Lexical	1278196	$< .001$	.25
N-Gram Word/POS	PCFG Syntactic	1824790	$< .001$	.73
N-Gram Word/POS	PCFG POS	1724892	$< .001$	.64
N-Gram Word/POS	RNNG	1081320	$< .001$	.08
N-Gram Word/POS	Transformer	1058508	.023	.06
PCFG Total	PCFG Lexical	1260934	$< .001$	.23
PCFG Total	PCFG Syntactic	1894848	$< .001$	.79
PCFG Total	PCFG POS	1788116	$< .001$	.69
PCFG Total	RNNG	1041841	.175	.04
PCFG Total	Transformer	1018956	.985	.02
PCFG Lexical	PCFG Syntactic	1653328	$< .001$	.58
PCFG Lexical	PCFG POS	1516045	$< .001$	.46
PCFG Lexical	RNNG	786684	$< .001$	.18
PCFG Lexical	Transformer	773722	$< .001$	.19
PCFG Syntactic	PCFG POS	679616	$< .001$	.27
PCFG Syntactic	RNNG	181541	$< .001$	.71
PCFG Syntactic	Transformer	195108	$< .001$	.70
PCFG POS	RNNG	294300	$< .001$	.61
PCFG POS	Transformer	303060	$< .001$	.60
RNNG	Transformer	973081	.985	.02

All groups have $n = 1409$.

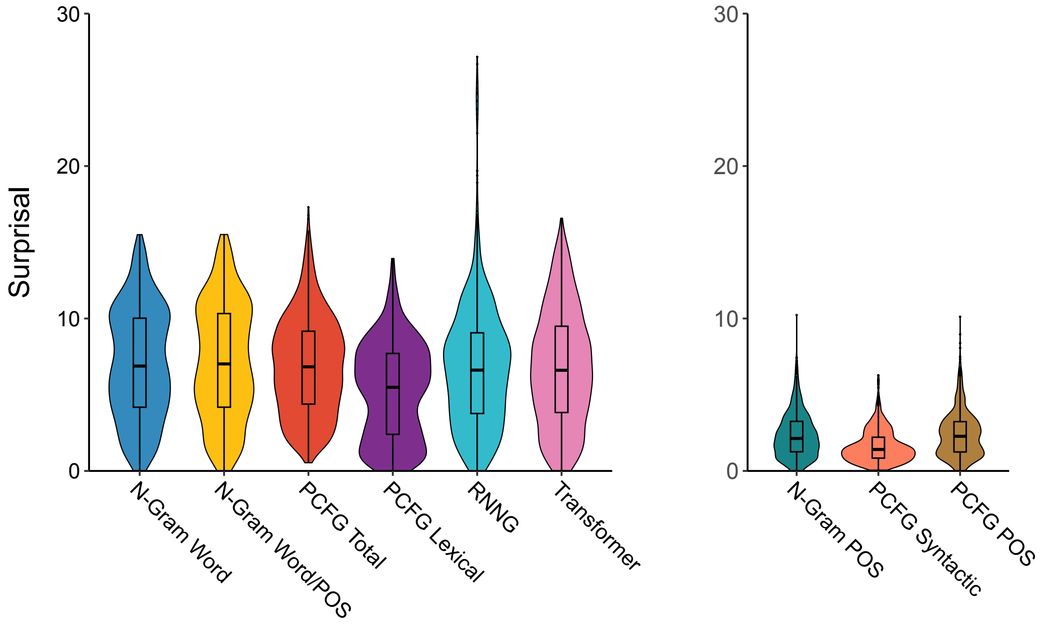


Figure 5.3: Distributions of stimuli surprisals produced by the language models.

5.5 Predicting Reading Times

5.5.1 Model Checking

The residuals of a GAMM should be inspected to check that model assumptions are not violated. The residuals should be approximately normally distributed. Inspection of Q-Q plots, histograms, and scatter plots showed that the residuals were approximately normally distributed for all GAMMs. Appendix B shows the plots for the base model for each participant group. Plots for the full models were omitted because they were visually indistinguishable from those for the base models for each language background.

5.5.2 Significance of Smooth Terms

Figure 5.4 shows the significance of the smooth and linear terms of the GAMMs developed for each surprisal type and participant group. The word order (SVO or SOV) and branching tendency (left or right) of each language are indicated on the figure. These disinctions will be relevant in our discussion. Note that s() is used to denote smooth terms fitted with thin-plate regression splines while ti() is used to denote tensor product interaction terms. A covariate's smooth term is significant if there is an association between that covariate and reading time. The complete statistical information for the models can be found in Appendix C. The participant, trial, word, and word position smooths were significant across all models for all participant groups. That is, there was an association between these covariates and reading time regardless of the reader's L1 and what surprisal type is included in the model. Measures of English language proficiency were not significant for the GAMMs developed using data from Chinese and Korean native speakers. In contrast, adjusted performance on the vocabulary test was a significant predictor of reading time for English native speakers for all models while reading comprehension adjusted performance was significant for Spanish native speakers for all models. Word length was generally a significant predictor of reading time for all participant groups, but especially for Chinese native speakers for whom it was significant across all models.

At least one of the terms including surprisal were significant for all models for all participant groups. Which surprisal term was significant differed between the reader's L1 and surprisal type. Some trends emerged. In particular, the sentence position and surprisal interaction term was significant for all surprisals derived from language models that explicitly used lexical information for all participant groups, with the exception of the RNNG surprisal for Chinese native speakers. Furthermore, this interaction term was non-significant for surprisals computed using only syntactic information, with the exception of the PCFG POS surprisal for Spanish native speakers. For English native speakers, it was also observed that the word position and surprisal interaction term was significant across all surprisal types except for PCFG POS surprisal.

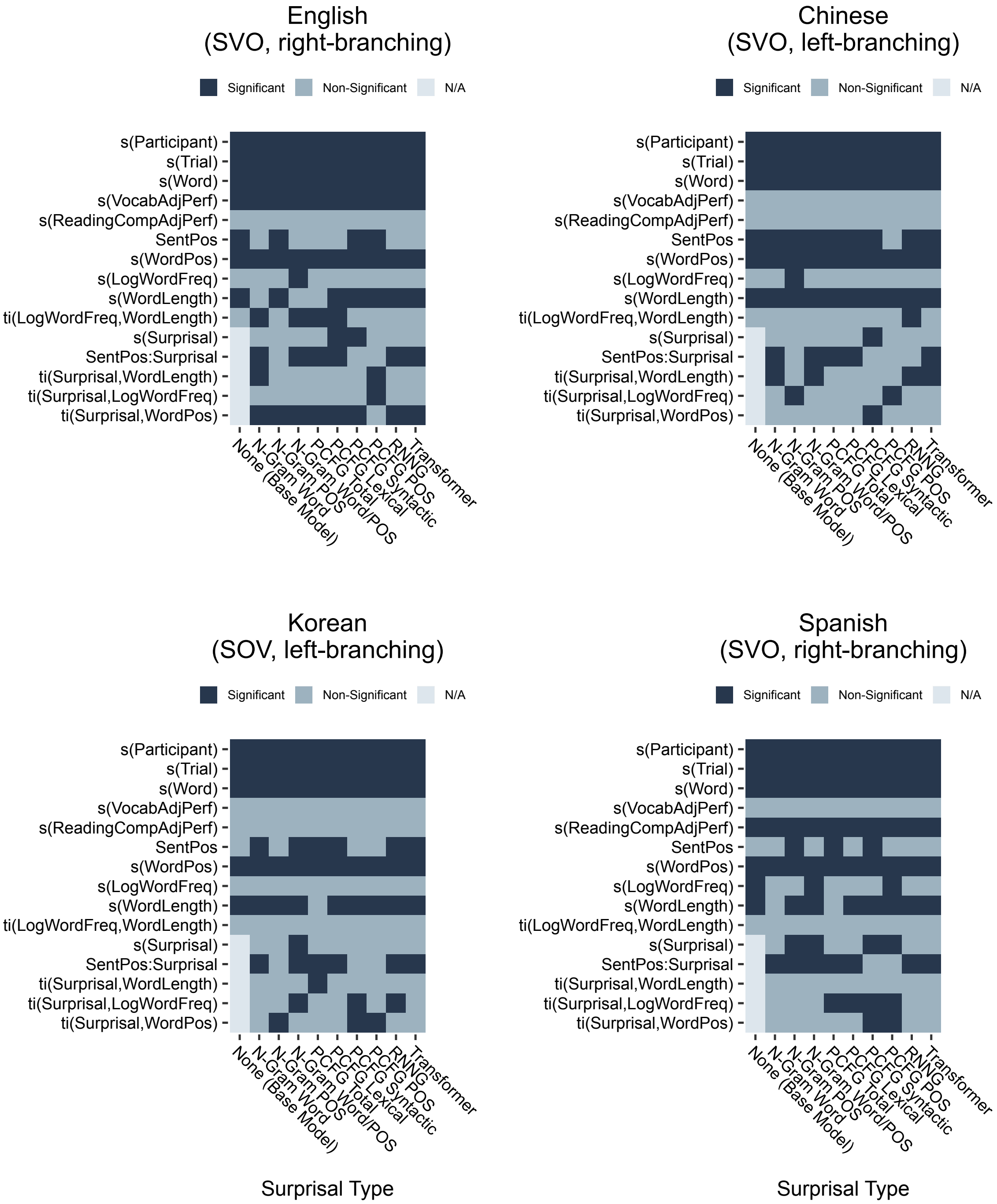


Figure 5.4: Term significance for each GAMM by first language.

5.5.3 Model Comparison

Figure 5.5 shows the ΔAIC values for each participant group. The word order (SVO or SOV) and branching tendency (left or right) of each language are indicated on the figure. These disinctions will be relevant in our discussion. Higher values of ΔAIC

indicate better full model fit to the data in terms of capacity for prediction. The surprisals are organized from highest to lowest ΔAIC for each participant group. This organization was chosen to draw focus to the difference in ΔAIC values within each participant group and discourage comparison of ΔAIC values amongst participant groups. This is important since AIC values should not be compared amongst models that were trained on different data.

Following a guideline, we considered that a $\Delta AIC < 2$ meant that the full model and base model were indistinguishable (Burnham & Anderson, 2003). That is, the addition of the surprisal terms did not improve the prediction of reading time to an extent that the added terms were justified. Based on this guideline, n-gram POS surprisal did not improve the base model for Korean native speakers. All other surprisals for all other first languages improved the base model for predicting reading time.

The best-performing surprisal for predicting reading times of English, Chinese, and Korean native speakers was the n-gram word surprisal, whereas for Spanish native speakers it was the n-gram word/POS surprisal. N-gram word surprisal and n-gram word/POS surprisal occupied two of the three highest-ranking positions for all participant groups. The remaining position in the top three was filled by a different surprisal for participants of each first language. For English native speakers, it was the PCFG POS surprisal, while, for Spanish native speakers, it was the PCFG syntactic surprisal. These two surprisal types both represent syntactic information in a hierarchical way. For Korean native speakers, it was the RNNG surprisal, which performed well especially in comparison to its relative performance for the other participant groups. For Chinese native speakers, the third best-performing surprisal was the transformer surprisal. Generally speaking, the PCFG lexical surprisal performed more poorly than the PCFG syntactic surprisal across all participant groups.

Several surprisal types performed essentially the same for Korean native speakers. These included all variations of the PCFG as well as the n-gram word/POS surprisal.

This contrasts the trend for other languages where the performance of most surprisals was distinguishable by ΔAIC.

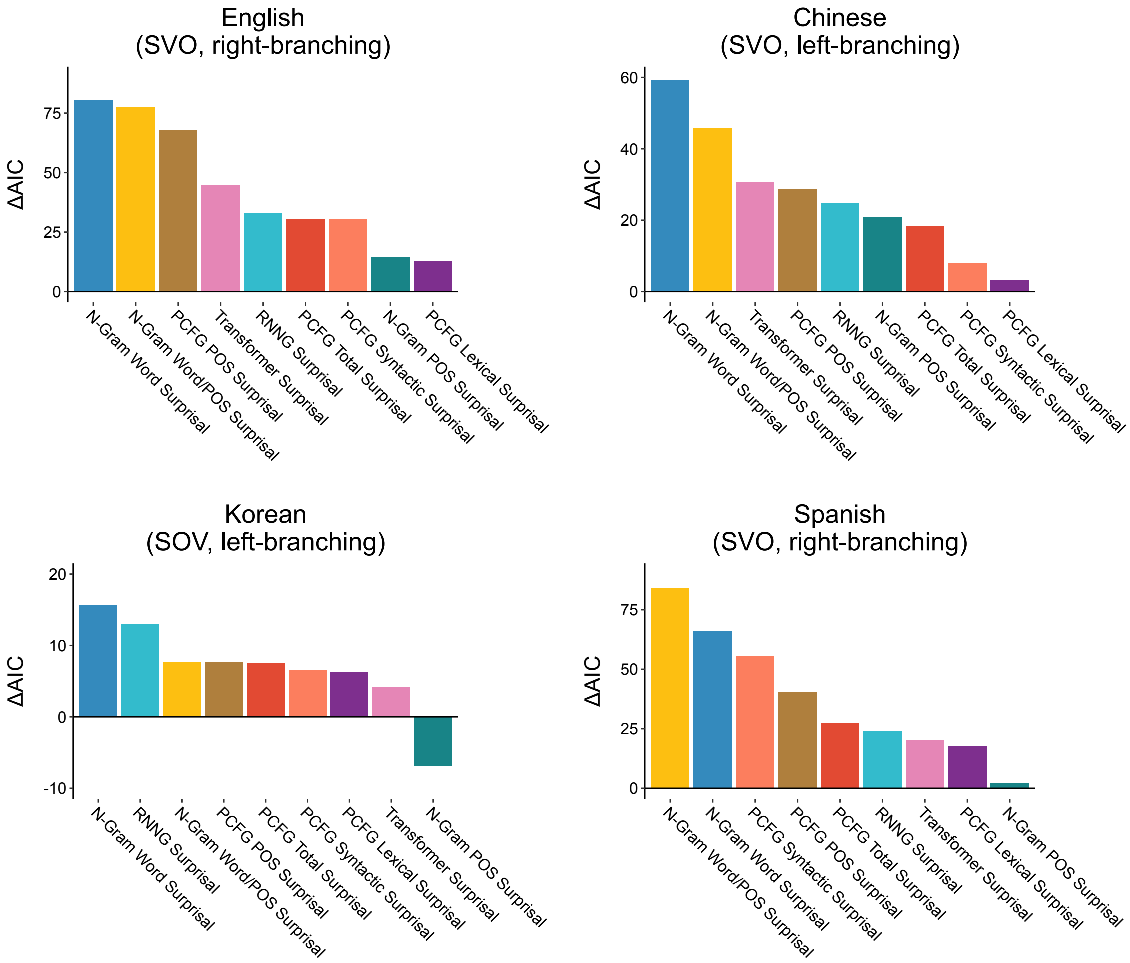


Figure 5.5: ΔAIC for each GAMM by first language.

Chapter 6

Discussion

Table 6.1 classifies the different surprisals by language model and training data. This terminology will be used to facilitate discussion throughout this section.

6.1 Differences in Surprisal Distributions by Language Model

Let us first consider the total surprisals. Our results showed that the distributions of n-gram word and n-gram word/POS surprisals differed from RNNG surprisals. This was to be expected considering that n-grams and RNNGs differ in both training data (sequential vs. hierarchical) and language model type (statistical vs. neural). The n-gram word/POS surprisals also differed from the transformer surprisals, but the n-gram word surprisals did not. It is reasonable that the n-gram word and transformer surprisals were not significantly different because they share the same sequential training data. When considering total surprisal, we saw that the PCFG surprisals were somewhere between those produced by an n-gram and those produced by a neural model. This was evidenced by the lack of significance in pairwise difference between the PCFG total surprisal and the n-gram word, n-gram word/POS, RNNG, and transformer surprisals. We also observed that adding explicit syntactic information when training the n-gram model did not significantly alter the distribution of surprisals produced, as shown by the lack of statistical difference between the n-gram word and

Table 6.1: Classification of surprisals.

Language Model	**Surprisal Type**	**Training Data**	**Surprisal Classification**
N-gram	N-gram Word	Word	Total
	N-gram POS	POS	Syntactic
	N-gram Word/POS	Word/POS	Total
PCFG	PCFG Total	Lexicalized Trees	Total
	PCFG Lexical	Lexicalized Trees	Lexical
	PCFG Syntactic	Lexicalized Trees	Syntactic
	PCFG POS	Unlexicalized Trees	Syntactic
RNNG	RNNG	Lexicalized Trees	Total
Transformer	Transformer	Word	Total

n-gram word/POS surprisals. This result was to be expected, considering our stimuli set had just 126 sentences and was not representative of English as a whole. In particular, 70% of the words in our stimuli occurred only once. As such, these words only occurred as one POS so their word representations were essentially equivalent to their word/POS representations. Additionally, the words that do occur more frequently in the stimuli were mostly words that only have one POS (e.g., the is always a determiner). Due to this, a large majority of word representations were the same in the word model and the word/POS model. As such, we cannot be certain that the distributions of n-gram word surprisals and n-gram word/POS surprisals would be similar for all corpora, especially for ones where words occur more frequently and as different parts of speech.

Now let us consider the syntactic surprisals. We saw that the n-gram and PCFG models that computed syntactic surprisal by using POS tags in place of words when training produced similar surprisal values. The PCFG syntactic surprisal values were distinct from these which may be explained by the fact that preceding lexical information was used to compute the syntactic surprisal of the current word. In contrast, the n-gram POS surprisals and PCFG POS surprisals were not informed by any lexical information. All syntactic surprisals were statistically different from all total and lexical surprisals, suggesting that they reflect different linguistic information.

6.2 Commonalities in Significance of GAMM terms across First Languages

The random effects for participant, trial, and word as well as the fixed effect for word position were significant predictors of reading time across first languages and surprisal types. This makes sense for the included random effects since the individual differences among participants, familiarity and fatigue with the task, and what word was read would be expected to be significant regardless of language background. Aside from the random effects, word position was the only fixed effect that was consistently significant across all language backgrounds and surprisal types. Previous research has found that word position has an effect on reading time independent of lexical and contextual predictors, such as word predictability, for L1 reading in English, Dutch, and German (Kuperman et al., 2010). Our results provide further evidence that the effect of word position on reading time is independent of other predictors, including surprisal derived from various language models, since word position is significant for all GAMMs. Considering that this is the case across all four language backgrounds, it may also suggest that this observation holds more broadly for both L1 and L2 reading in English.

6.3 Commonalities in Surprisal Performance across First Languages

The best-performing surprisals for all participant groups were derived from variations of n-gram language models that included lexical information along with either implicit or explicit syntactic information. The strong performance of n-grams here contrasts previous research findings that language models with low perplexity tend to produce surprisals that better predict reading times (Goodkind & Bicknell, 2018). As seen in Table 5.4, the n-gram word language model had a higher perplexity on the test set than the transformer, so it would have been expected that the transformer would perform better for predicting reading times. Other research suggests that there is an upper bound to the correlation between perplexity and predictive power of surprisal for reading times with very large language models being outperformed by some of their smaller counterparts that have higher perplexities (de Varda & Marelli, 2023; Shain et al., 2022). There are differences among these studies and ours in terms of the datasets used to train the language models, the type of reading-time data used (e.g., eye-tracking or self-paced), and the covariates included in the predictive model of reading time. Considering these methodological differences and the differences in results, it seems that the sometimes observed trend of lower perplexity language models producing surprisals that better predict reading times does not hold generally.

Our results showed that PCFG lexical surprisal performed more poorly than PCFG syntactic and PCFG POS surprisal for predicting reading time, regardless of the readers' first language. This seems to indicate that lexical information separated from syntactic is not particularly good at predicting reading times. Furthermore, the n-gram POS surprisal performed more poorly than the PCFG POS surprisal across all participant groups, which seems to suggest that capturing the hierarchical aspect of syntax is important when syntactic information is separated from lexical information. However, the top-two-performing surprisals for all participant groups used a combination of lexical and syntactic information. This makes sense since both lexical and syntactic knowledge are important for reading and there are interactions between them for both L1 and L2 reading (Barnett, 1986).

6.4 Differences in Surprisal Performance by First Language

Table 6.2 compares the top-performing surprisals for L1 speakers of left-branching and right-branching languages. Both of the right-branching languages have a syntactic surprisal derived using the full preceding context with hierarchical data performing in the top three. Conversely, surprisals computed using the full preceding context with both syntactic and lexical information are present in the top three for the left-branching languages. Amici et al. (2019) found that word order of first language can predict a person's working memory, with native speakers of left-branching languages able to recall initial stimuli more easily. It is possible that this observation is relevant to our findings. Due to higher working memory capacity, perhaps native speakers of left-branching languages use more context when predicting upcoming words during reading. In this case, it would make sense that surprisals computed using richer context performed more strongly for speakers of left-branching languages since the surprisals would better approximate the reader's predictions.

Table 6.2: Comparison of surprisal performance for left-branching and right-branching languages.

	Left-branching	**Right-branching**
Languages	Chinese Korean	English Spanish
Commonalities in Top 3	N-gram Word N-gram Word/POS	N-gram Word N-gram Word/POS
Differences in Top 3	Transformer RNNG	PCFG POS PCFG Syntactic
	Both use the full context with syntactic and lexical information	Both use the full context with only syntactic information

Although English and Spanish are both right-branching SVO languages, word order in Spanish is somewhat more flexible than in English. In particular, the placement of adjectives is more flexible and the typical SVO word order of Spanish is more frequently modified to change the emphasis of a sentence (Butt & Benjamin, 1994). This difference might help explain why different versions of syntactic surprisal are in the top performers for English and Spanish. For English, the PCFG POS surprisal was the third best performer while, for Spanish, it was the PCFG syntactic surprisal. These two surprisals are similar in many ways, reflecting the similarities of English and Spanish. However, PCFG POS surprisal used only syntactic information from the preceding context and current word to compute surprisal. In contrast, the PCFG syntactic surprisal used both syntactic and lexical information from the preceding context but only syntactic information from the current word. In this way, the PCFG syntactic surprisal uses a richer context. Perhaps this richer context is important for predicting English reading times of L1 speakers of languages that have more flexible word orders than English. Chinese and Korean also have more word order flexibility (Kim, 2018; Sun & Givon, 1985), and the prediction of English reading times for L1 speakers of these languages is also better supported by surprisals computed using richer context.

Chinese and Korean are both left branching but Chinese has SVO word order while Korean has SOV. In this way, Korean is more syntactically distinct from English than Chinese is. Perhaps this could explain why the RNNG surprisal was in the top three for Korean native speakers while the transformer surprisal was in the top three for Chinese native speakers. The RNNG uses explicit hierarchical syntactic information when computing surprisal whereas the transformer only has access to sequential information. Furthermore, the performance of all PCFG surprisals, which were informed by explicit syntactic information, was also strong for Korean native speakers (moreso than for Chinese and Spanish native speakers). It could be the case that explicit syntactic information is important for predicting the English reading times of people whose first languages have a different word order than English.

6.5 Limitations

Computations of surprisal are dependent on the data used to train the language model. Our language models were trained on the WikiText-2 dataset, a corpus of Wikipedia articles. This training data was appropriate for our use case since the articles were written in a style of English similar to the passages participants read from the New York Times. However, Wikipedia articles do not cover the wide variety of genres that an English speaker would typically be exposed to. As such, surprisals computed from the WikiText-2 corpus likely differ from a person's internal representation of word predictability. Using a larger corpus for training that includes more genres would likely help the computed surprisal values better approximate human-like expectation.

Our results are also tied to the implementations of each language model that we used to compute surprisals. It is possible that there could be undiscovered errors or bugs in the implementations we used. In such a case, our findings could change if we were to use different implementations.

Our results are based on a sample of L1 speakers of English, Chinese, Korean, and Spanish. Participants had somewhat varied levels of English language proficiency, but there was little variety in participant age and gender. Although English language skills varied by participant, they still represented a certain ability range and beginning learners of English were not represented. Accordingly, making generalizations to the complete populations of L1 speakers from these language backgrounds is not possible.

Our work involved English speakers with different language backgrounds. The four different first languages of the participants have differences in writing system, word order, and branching. However, there are over 7,000 living languages spoken across the world, each unique in some way (Ebhherhard et al., 2023). Accordingly, our results cannot be generalized to all L1 speakers of SOV languages, for example, based on the results for L1 speakers of Korean alone. However, we have studied widely spoken languages; there are 1.35 billion L1 speakers of Chinese, 485 million L1 speakers of Spanish, and 82 million L1 speakers of Korean (Ebherhard et al., 2023). As such, our work is relevant for a large number of people.

6.6 Ethical Considerations

Participants in our study read texts written in American standard English. Participants may have varying degrees of familiarity with American standard English since many different varieties of English exist around the world. Our work involved participants who had lived in the U.S. for some time so the use of American standard English was justifiable. However, future work involving L2 English speakers with diverse language backgrounds should also consider the diversity of English spoken around the world.

6.7 Future Work

Future work could include a wider variety of language backgrounds. Studying several SVO languages and SOV languages could allow for stronger conclusions to be drawn about the relationship between L1 word order and the surprisals that best predict L2 reading times in English. Another avenue to pursue would be training the language models on a larger and more varied corpus to better estimate surprisal. Future studies could also explore the relationship between English proficiency and the surprisals that best predict reading times. This work also presents the opportunity to better understand large language models that operate effectively as black boxes by evaluating their performance for predicting human reading times.

Chapter 7

Conclusion

Prediction is key in L1 reading (Huettig, 2015; Kuperberg & Jaeger, 2016; F. Smith, 1975). This claim is supported by the ability of surprisal derived from language models to predict reading time (Goodkind & Bicknell, 2018; Hale, 2001; Hale et al., 2018; N. J. Smith & Levy, 2013). Prediction also appears to play a role in L2 reading (Berzak & Levy, 2022; Chun, 2020) but perhaps to a lesser extent (Grüter et al., 2014). Attention is now being directed to exploring the utility of surprisal for predicting L2 reading times (de Varda & Marelli, 2022). However, much is still unknown about the relationship between language background and the performance of surprisal for predicting L2 reading times.

To narrow this gap in knowledge, we trained nine language models that varied in the extent of syntactic information, lexical information, and preceding context they used to compute surprisal. We developed GAMMs to predict reading times of English speakers with various first languages (English, Chinese, Korean, and Spanish) using these different surprisals. For each participant group, we compared the performance of the different surprisals for predicting reading time using AIC.

We found several similarities in performance of the different surprisals for predicting English reading times across language backgrounds. Surprisals derived from a standard n-gram and an n-gram with added POS tags were among the top three performers for all language backgrounds. For PCFG surprisals, the lexical portion of the total surprisal performed worse than the syntactic portion. For the syntactic surprisals, those derived from language models trained on hierarchical data performed better than those trained on sequences.

Apart from these commonalities, we observed several differences. We found that the prediction of English reading times of L1 speakers of left-branching languages benefited from the inclusion of surprisal with richer context. We also found that surprisals computed using hierarchical syntactic information performed better for L1 speakers of Korean (an SOV language) as opposed to L1 speakers of languages with the same word order as English (SVO). Further research involving more language backgrounds and participants is needed to better understand these differences.

Our work furthers the emerging research on using language model-derived surprisal for predicting L2 reading times. Our findings contribute to this area by showing that surprisals computed by different language models perform differently for predicting English reading times based on the language background of the reader. Our work indicates that a one-size-fits-all approach would leave some groups with sub-optimal performance for L2 English reading time prediction. As such, it is important to consider language background when using language models in the study of L2 reading.

Bibliography

AlJassmi, M. A., Warrington, K. L., McGowan, V. A., White, S. J., & Paterson, K. B. (2022). Effects of word predictability on eye movements during Arabic reading. *Attention, Perception, & Psychophysics*, *84*(1), 10–24. https://doi.org/10.3758/s13414-021-02375-1

Amici, F., Sánchez-Amaro, A., Sebastián-Enesco, C., Cacchione, T., Allritz, M., Salazar-Bonet, J., & Rossano, F. (2019). The word order of languages predicts native speakers' working memory. *Scientific Reports*, *9*, Article 1124. https://doi.org/10.1038/s41598-018-37654-9

Anatole, L. (1997). *An introduction to the languages of the world*. Oxford University Press.

Barnett, M. A. (1986). Syntactic and lexical/semantic skill in foreign language reading: Importance and interaction. *The Modern Language Journal*, *70*(4), 343–349. https://doi.org/https://www.jstor.org/stable/326811

Berzak, Y., & Levy, R. P. (2022). *Eye movement traces of linguistic knowledge* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/mw2gv

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Booth, T. L., & Thompson, R. A. (1973). Applying probability measures to abstract languages. *IEEE Transactions on Computers*, *C-22*(5), 442–450. https://doi.org/10.1109/T-C.1973.223746

Brown, J. I., Fishco, V. V., & Hanna, G. (1993). *Nelson-Denny reading test*. The Riverside Publishing Company.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer, New York.

Butt, J., & Benjamin, C. (1994). Word order. In *A New Reference Grammar of Modern Spanish* (pp. 464–475). Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-8368-4_37

Chun, E. (2020). L2 prediction guided by linguistic experience. *English Teaching*, *75*, 79–103. https://doi.org/10.15858/engtea.75.s1.202006.79

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Pyschological Review*, *108*(1), 204–256.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

de Varda, A. G., & Marelli, M. (2022). The effects of surprisal across languages: Results from native and non-native reading. *Findings of the Association for Computational Linguistics: AACL-IJCNLP*, 138–144.

de Varda, A. G., & Marelli, M. (2023). Scaling in cognitive modelling: A multilingual approach to human reading times. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers*, 139–149.

Dryer, M. S. (2013). Order of subject, object and verb. In M. S. Dryer & M. Haspelmath (Eds.), *WALS Online (v2020.3) [Data set]*. Zenodo. https://doi.org/10.5281/zenodo.7385533

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 199–209. https://doi.org/10.18653/v1/N16-1024

Ebherhard, D. M., Simons, G. F., & Fennig, C. D. (2023). *Ethnologue: Languages of the world* (26th ed.). Dallas, Texas: SIL International. http://www.ethnologue.com

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

Fernández, G., Shalom, D. E., Kliegl, R., & Sigman, M. (2014). Eye movements during reading proverbs and regular sentences: The incoming word predictability effect. *Language, Cognition and Neuroscience*, *29*(3), 260–273. https://doi.org/10.1080/01690965.2012.760745

Gass, S. (1979). Language transfer and universal grammatical relations. *Language Learning*, *29*(2), 327–344. https://doi.org/10.1111/j.1467-1770.1979.tb01073.x

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 10–18. https://doi.org/10.18653/v1/W18-0102

Grüter, T., Rohde, H., & Schafer, A. J. (2014). The role of discourse-level expectations in non-native speakers' referential choices. *BUCLD 38: Proceedings of the 38th annual Boston University Conference on Language Development.*

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *NAACL '01: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, 1–8. https://doi.org/10.3115/1073336.1073357

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. (2018). Finding syntax in human encephalography with beam search. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2727–2736. https://doi.org/10.18653/v1/P18-1254

Hammarström, H. (2016). Linguistic diversity and language evolution. *Journal of Language Evolution*, *1*(1), 19–29. https://doi.org/10.1093/jole/lzw002

Heafield, K. (2011). KenLM: Faster and smaller language model queries. *Proceedings of the 6th Workshop on Statistical Machine Translation*, 187–197. https://aclanthology.org/W11-2123/

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034

Kim, N. (2018). Korean. In *The world's major languages* (pp. 781–796). Routledge.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 423–430. https://doi.org/10.3115/1075096.1075150

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262–284. https://doi.org/10.1080/09541440340000213

Koda, K. (1990). The use of L1 reading strategies in L2 reading: Effects of L1 orthographic structures on L2 phonological recoding strategies. *Studies in Second Language Acquisition*, *12*(4), 393–410. https://doi.org/10.1017/S0272263100009499

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology*, *63*(9), 1838–1857. https://doi.org/10.1080/17470211003602412

Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Lõo, K., Marelli, M., … Usal, K. A. (2023). Text reading in English as a second language: Evidence from the Multilingual Eye-movements Corpus. *Studies in Second Language Acquisition*, *45*(1), 3–37. https://doi.org/10.1017/S0272263121000954

Law, S., Wong, W., & Kong, A. (2006). Direct access from meaning to orthography in Chinese: A case study of superior written to oral naming. *Aphasiology*, *20*(6), 565–578. https://doi.org/10.1080/02687030600591799

Li, X., Bicknell, K., Liu, P., Wei, W., & Rayner, K. (2014). Reading is fundamentally similar across disparate writing systems: A systematic characterization of how words and characters influence eye movements in Chinese reading. *Journal of*

Experimental Psychology: General, *143*(2), 895–913. https://doi.org/10.1037/a0033580

Li, X., Huang, L., Yao, P., & Hyönä, J. (2022). Universal and specific reading mechanisms across different writing systems. *Nature Reviews Psychology*, *1*(3), 133–144. https://doi.org/10.1038/s44159-022-00022-6

Li, X., Liu, P., & Rayner, K. (2011). Eye movement guidance in Chinese reading: Is there a preferred viewing location? *Vision Research*, *51*(10), 1146–1156. https://doi.org/10.1016/j.visres.2011.03.004

Li, X., & Pollatsek, A. (2020). An integrated model of word processing and eye-movement control during Chinese reading. *Psychological Review*, *127*(6), 1139–1162. https://doi.org/10.1037/rev0000248

Luong, M., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. https://doi.org/10.18653/v1/D15-1166

McNeill, D., Yukawa, R., & McNeill, N. B. (1971). The acquisition of direct and indirect objects in Japanese. *Child Development*, *42*, 237–249.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2017). Pointer sentinel mixture models. *5th International Conference on Learning Representations.*

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. https://doi.org/10.18653/v1/2021.cmcl-1.2

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 398–408.

Noji, H., & Oseki, Y. (2021). Effective batching for recurrent neural network grammars. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4340–4352. https://doi.org/10.18653/v1/2021.findings-acl.380

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception*, *8*(1), 21–30. https://doi.org/10.1068/p080021

Rayner, K., Li, X., Juhasz, B. J., & Yan, G. (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review*, *12*(6), 1089–1093. https://doi.org/10.3758/BF03206448

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504–509. https://doi.org/10.3758/BF03214555

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *27*(2), 249–276. https://doi.org/10.1162/089120101750300526

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, *1*, 324–333. https://doi.org/10.3115/1699510.1699553

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. P. (2022). *Large-scale evidence for logarithmic effects of word predictability on reading time* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/4hyna

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Kwon, N., Lõo, K., . . . Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, *54*(6), 2843–2863. https://doi.org/10.3758/s13428-021-01772-6

Smith, F. (1975). The role of prediction in reading. *Elementary English*, *52*(3), 305–311. https://www.jstor.org/stable/41592609

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Stern, M., Fried, D., & Klein, D. (2017). Effective inference for generative neural parsing. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1695–1700. https://doi.org/10.18653/v1/D17-1178

Sun, C., & Givon, T. (1985). On the so-called SOV word order in Mandarin Chinese: A quantified text study and its implications. *Language*, *61*(2), 329–351. https://doi.org/10.2307/414148

Tan, L. H., & Perfetti, C. A. (1997). Visual Chinese character recognition: Does phonological information mediate access to meaning? *Journal of Memory and Language*, *37*(1), 41–57. https://doi.org/10.1006/jmla.1997.2508

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*(2), 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x

van Kooten, P. (2022). Contractions. https://pypi.org/project/contractions/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *NIPS '17: Proceedings of the 31st Conference on Neural Information Processing Systems*.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686. https://doi.org/10.1198/016214504000000980

Wood, S. (2006). *Generalized additive models: An introduction with R*. CRC Press.

Woolley, G. (2011). *Reading comprehension: Assisting children with learning difficulties*. Springer, Netherlands. https://doi.org/10.1007/978-94-007-1174-7

Yun, H., Lee, D., Nam, Y., & Hong, U. (2017). The predictability effect on eye movement in reading Korean dative sentences. *Language and Information*, *21*(1), 73–99. https://doi.org/10.29403/LI.21.1.4

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832

Appendix A: Stimuli

Here is a complete list of the stimuli used in the study.

- Very few issues can bring together lawmakers of both parties. Animal cruelty is one of them.
- Koalas have been running into hard times. They have suffered for years from habitat destruction, dog attacks, automobile accidents.
- Pancreatic cancer has a bad reputation. It is a terrible disease, but most people do not realize there are ways that early detection can help.
- Some nutrition experts say eggs are good for you, even though they are high in cholesterol. Others are sure they are bad.
- Experts generally agree that watching a lot of television is bad for children. Now a new study suggests it may not be very good for adults, either.
- Legislators and gun safety advocates often focus on how guns are purchased. But many lives could be saved, especially among children, if they looked more at how they are stored.
- In general, any exercise improves our health. But a wealth of recent science and personal experience indicate that different people can respond to similar exercise routines in different ways.
- When new breast cancer drugs come to market, there is little data to indicate whether they are safe or effective in men. Some new drugs are approved only for women.
- An extensive study confirms that red meat might not be that bad for you. But it is bad for the planet, with chicken and pork less harmful than beef.

- The ground around the building slopes away. That allows the classroom windows to be low enough for children to see out while still being high above the ground.

- Farmers and plant breeders are in a race against time. The world population is growing rapidly, requiring ever more food, but the amount of cultivable land is limited.

- One benefit of discovery is that you get to name the things you discovered. Astronomy is blessed in this regard.

- The overwhelming majority of snails live in shells that coil to the right. But occasionally some are born with shells that turn the other way.

- California has positioned itself as a leader in the fight against climate change. It leads the nation by far in solar power.

- Climate change will inevitably transform the way the world produces goods. Farmers who produce wine grapes, an especially sensitive crop, are already feeling those effects.

- The many whiskeys of Japan are complex and often expensive. Learning about them can be a challenge.

- Great snow conditions can make a ski trip magical. Skiers can weight the odds in their favor by picking their destinations according to the time of winter.

- Earth is warming, and we know why. Light is reflected and absorbed by clouds, air, oceans, ice and land.

- Nobody ever really tells you how to breathe. You just know how to do it.

- Imagine an animal that looks like a dinosaur, and you probably will not imagine a bat. But that may change.

- Video games work hard to hook players. Designers use predictive algorithms and principles of behavioral economics to keep fans engaged.

- Americans are feeling better about the economy. Or at least republicans are.

- Parasites are a huge force in the natural world. For the most part they simply feed on their hosts.

- There should be no question that autistic people need and value social connections. We all do.
- Once upon a time, the sky was everything above the earth, which lay at the center of human experience. The sky was the place of gods.
- The abundance of uncontrolled case studies can give an impression that a single dietary item, like juice, is linked with obesity. But they cannot determine cause and effect.
- Americans are having fewer babies. At first, researchers thought the declining fertility rate was because of the recession, but it kept falling even as the economy recovered.
- New research suggests that more money really does lead to a more satisfying life. Surveys of thousands of Swedish lottery winners have provided persuasive evidence of this truth.
- Milan has no sea. It has no river or lake, just a few ebbing canals.
- Morality exists only because we evolved the capacity to empathize. And empathy provides the best guide to moral action.
- Story creates culture. It teaches us to feel, think and behave in ways generally approved of by those around us.
- Hospitals are often thought of as the hubs of our health care system. But hospital closings are rising, particularly in some types of communities.
- Two sexes have never been enough to describe human variety. Not in biblical times and not now.
- The biggest racial preferences in this country have nothing to do with college admissions or job offers. They have to do with political power.
- Businesses can deliver everything on demand, from dinner to dry cleaning. Some will even show up at your door to give you cupcakes or walk your dog.
- Planning for vacation is fun. We make lists of clothes to pack and museums to visit, anticipating the relaxation and fun to come.
- Certain types of wine seem incapable of winning popular acceptance. Riesling is one, particularly in its moderately sweet form.

- Bankers kept their name tags obscured behind ties. Many tried to keep a low profile and avoided talking to the news media.

- Hollywood executives are not so different from the stars they employ. They want to be seen in the right booths at the right restaurants.

- The more we learn about how people really think, the more we must rethink economic theory. Changes in fundamental beliefs play a major role in the fluctuations of the economy.

- History suggests that the world is about due for another financial crisis. One of the places it might start, according to a growing number of indicators, is Italy.

- Few doubt that energy has improved lives and enabled human progress. Yet one of the biggest challenges facing the world is the polarized debate over the future of energy.

- Dance is about change. The body keeps altering its shape while we watch it move.

- Plants have no eyes, no ears, no mouth and no hands. They do not have a brain or a nervous system.

- There are many more gorillas and chimpanzees than previously believed, new research finds. Nonetheless, their numbers are rapidly declining.

- Forests are disappearing. Maps show shrinking woodlands all over the world.

- A cowboy walks into a bar. Or more specifically, a Nigerian immigrant grad student in a cowboy hat walks into a bar in Oklahoma.

- One man vanished diving at midnight. Another was attacked by a great white shark in deep water.

- Student loans are central to financing college educations, yet millions of borrowers are in default. That is clear evidence that the system is in dire need of improvement.

- People seem to be more worried than ever about stress. We hear that stress can lead to everything from depression to cancer.

- Music can transport us back in time or help usher in new ways of thinking. It can capture a historical moment or offer an escape from the world around us.

- The laws that authorize medical aid in dying have worked exactly as intended, the evidence shows. They have benefited dying patients and their families without causing any harm to anyone.

- Israel cannot define a Jew. And the more it attempts to do so, the more obvious that becomes.

- Corporate tax cuts will put billions of dollars back in the hands of businesses this year. Naturally, people want to know how those businesses will spend it.

- Florida, it seems, has always been a popular destination. Even the first known Americans gravitated to the state.

- When early humans discovered how to build fires, life became much easier in many regards. They huddled around fire for warmth, light and protection.

- For most of human history, nobody ate a vegetable for pleasure. After all, vegetables have no evolutionary imperative to be delicious.

- Animals around the world are on the move. So are their parasites.

- Gifts are often left on front row seats at fashion shows for guests. Rarely, however, are they placed on every seat.

- Trying to figure out why humans cry is exhausting. We cry about death, violence, breakups, abandoned puppies, sweet kisses and words charged with all kinds of meanings.

- The most popular movie in India this summer is about a toilet. It nearly causes a divorce.

- Bargain hunters have begun to pay close attention to emerging market stocks. They are a little late to the game.

- One new drug promises to stop cancer from spreading to other organs. Another would treat blood cancer.

Appendix B: Model Checking

Figures B.1–B.4 show four plots of the residuals for the GAMM base models for each participant group. The plots show that the residuals are approximately normally distributed with some deviation from normality at the tails of the distributions.

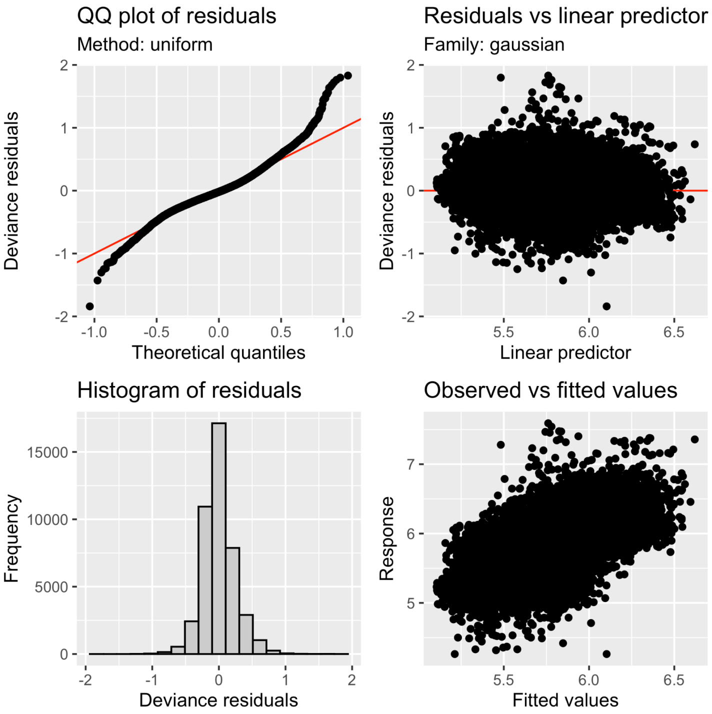


Figure B.1: Residual plots for the GAMM base model for English native speakers.

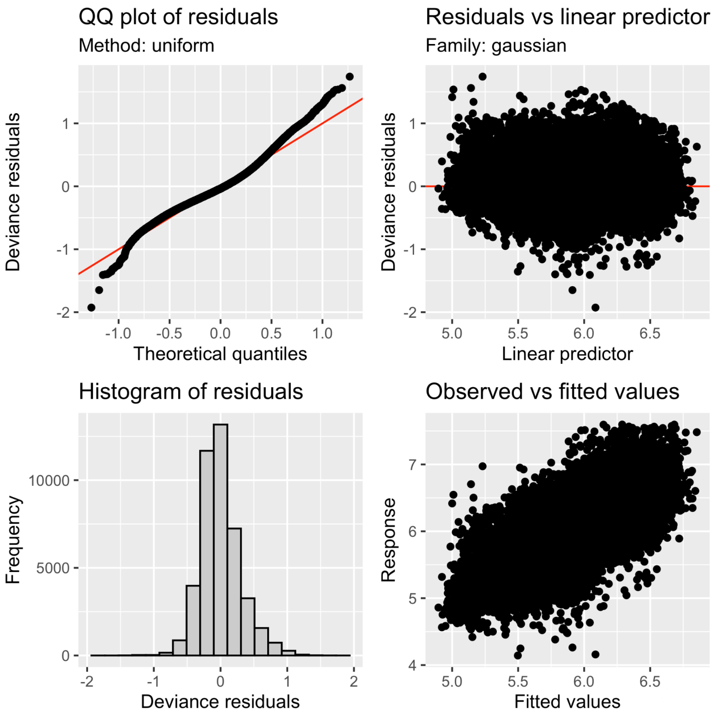


Figure B.2: Residual plots for the GAMM base model for Chinese native speakers.

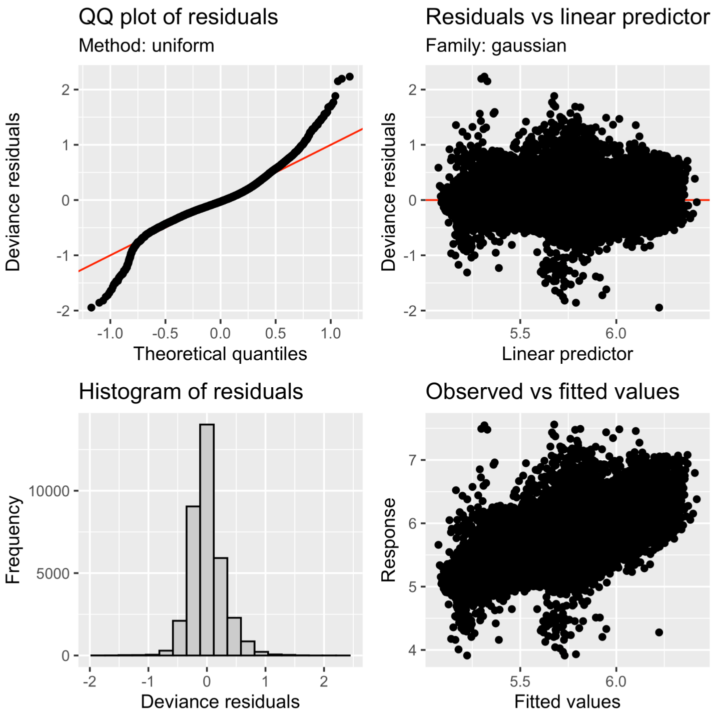


Figure B.3: Residual plots for the GAMM base model for Korean native speakers.

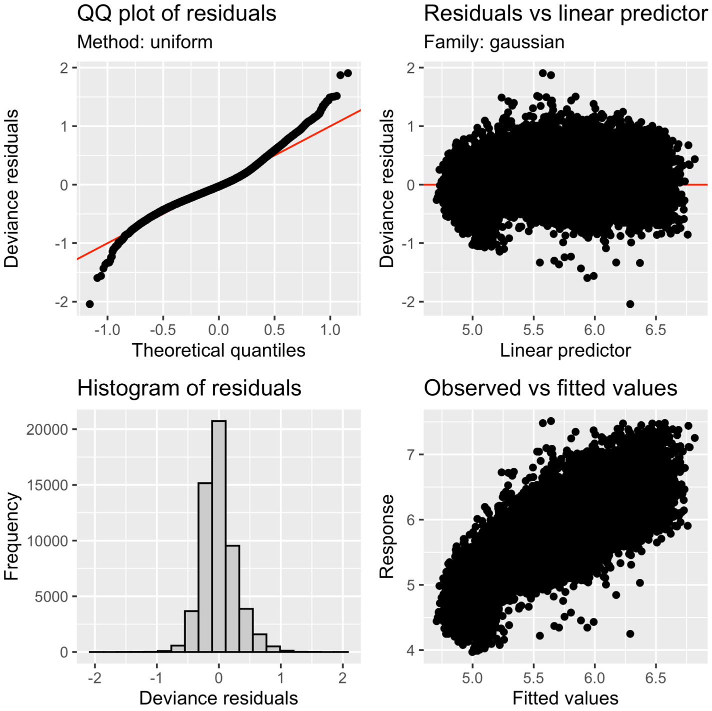


Figure B.4: Residual plots for the GAMM base model for Spanish native speakers.

Appendix C: GAMM Statistics

Tables C.1–C.10 provide statistical summaries of the terms in the GAMMs. Eff. df refers to effective degrees of freedom. Ref. df refers to reference degrees of freedom.

Table C.1: Statistical summary of the terms in the base models.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.741	0.035	165.07	< .001	5.907	0.050	118.83	< .001
SentPos	0.023	0.003	7.67	< .001	-0.033	0.004	-9.31	< .001
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1083.21	< .001	32.0	32.0	1724.68	< .001
s(Trial)	61.4	62.0	233.89	< .001	61.1	62.0	178.45	< .001
s(Word)	508.9	687.0	4.03	< .001	455.2	687.0	2.72	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.935	1.0	1.0	0.45	.505
s(WordPos)	7.9	8.5	16.77	< .001	6.3	7.2	52.77	< .001
s(LogWordFreq)	1.0	1.0	2.03	.154	1.0	1.0	3.00	.083
s(WordLength)	1.0	1.0	12.23	< .001	1.0	1.0	41.25	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	0.70	.403	3.5	3.9	2.18	.066
	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.717	0.053	107.87	< .001	5.771	0.050	115.65	< .001
SentPos	-0.006	0.004	-1.52	.128	0.009	0.003	3.05	.002
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1392.44	< .001	41.0	41.0	2783.95	< .001
s(Trial)	54.9	62.0	21.48	< .001	61.5	62.0	532.86	< .001
s(Word)	317.8	687.0	1.34	< .001	494.7	687.0	3.35	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.964	1.0	1.0	0.00	.980
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.94	.005
s(WordPos)	6.2	7.1	17.93	< .001	7.9	8.5	41.77	< .001
s(LogWordFreq)	1.0	1.0	0.19	.661	1.0	1.0	4.76	.029
s(WordLength)	1.0	1.0	22.34	< .001	1.0	1.0	18.79	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	0.02	.882	1.0	1.0	0.06	.810

Table C.2: Statistical summary of the terms in the full model with n-gram word surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.735	0.035	164.67	< .001	5.903	0.050	118.71	< .001
SentPos	0.006	0.006	1.06	.291	-0.049	0.007	-6.88	< .001
SentPos:NgramWordSurp	0.003	0.001	3.11	.002	0.002	0.001	2.41	.016
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1075.31	< .001	32.0	32.0	1724.65	< .001
s(Trial)	61.4	62.0	230.79	< .001	61.1	62.0	184.59	< .001
s(Word)	495.9	687.0	3.64	< .001	447.4	687.0	2.60	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.936	1.0	1.0	0.44	.505
s(WordPos)	7.9	8.5	17.96	< .001	6.1	7.0	51.71	< .001
s(LogWordFreq)	1.0	1.0	2.11	.146	1.0	1.0	0.06	.811
s(WordLength)	1.0	1.0	0.57	.451	1.0	1.0	30.92	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	13.25	< .001	2.5	2.9	1.62	.182
s(NgramWordSurp)	1.0	1.0	0.15	.698	1.0	1.0	0.05	.826
ti(NgramWordSurp,WordLength)	7.4	8.6	3.33	.001	3.7	3.9	7.85	< .001
ti(NgramWordSurp,LogWordFreq)	1.0	1.0	2.00	.157	1.0	1.0	0.60	.440
ti(NgramWordSurp,WordPos)	8.7	10.4	1.94	.018	1.0	1.0	0.00	1.000
	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.714	0.053	107.77	< .001	5.764	0.050	115.44	< .001
SentPos	-0.029	0.007	-3.87	< .001	-0.010	0.006	-1.72	.086
SentPos:NgramWordSurp	0.003	0.001	3.52	< .001	0.003	0.001	3.59	< .001
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1391.40	< .001	41.0	41.0	2803.52	< .001
s(Trial)	54.9	62.0	20.65	< .001	61.5	62.0	671.88	< .001
s(Word)	309.5	687.0	1.26	< .001	489.3	687.0	3.35	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.978
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.95	.005
s(WordPos)	6.1	7.0	18.04	< .001	7.9	8.5	41.97	< .001
s(LogWordFreq)	1.0	1.0	0.14	.710	1.0	1.0	3.24	.072
s(WordLength)	1.0	1.0	5.19	.023	1.0	1.0	1.81	.178
ti(LogWordFreq,WordLength)	1.0	1.0	0.07	.794	1.0	1.0	2.98	.084
s(NgramWordSurp)	1.0	1.0	3.41	.065	1.0	1.0	0.72	.398
ti(NgramWordSurp,WordLength)	2.9	3.3	2.09	.092	3.5	3.7	3.64	.004
ti(NgramWordSurp,LogWordFreq)	1.0	1.0	1.47	.225	1.0	1.0	2.89	.089
ti(NgramWordSurp,WordPos)	1.0	1.0	0.08	.778	3.6	4.7	0.61	.659

Table C.3: Statistical summary of the terms in the full model with n-gram POS surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.743	0.035	165.06	< .001	5.908	0.050	118.84	< .001
SentPos	0.026	0.006	4.60	< .001	-0.031	0.007	-4.50	< .001
SentPos:NgramPOSSurp	-0.002	0.002	-0.90	.369	-0.001	0.002	-0.51	.609
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1079.31	< .001	32.0	32.0	1722.91	< .001
s(Trial)	61.4	62.0	230.01	< .001	61.1	62.0	207.48	< .001
s(Word)	508.9	687.0	3.96	< .001	451.5	687.0	2.66	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.935	1.0	1.0	0.45	.505
s(WordPos)	8.0	8.5	16.38	< .001	6.5	7.3	48.35	< .001
s(LogWordFreq)	1.0	1.0	2.47	.116	1.0	1.0	4.12	.042
s(WordLength)	1.0	1.0	11.69	.001	1.0	1.0	35.50	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	0.71	.400	3.5	3.9	2.34	.059
s(NgramPOSSurp)	1.0	1.0	1.63	.202	1.0	1.0	1.47	.225
ti(NgramPOSSurp,WordLength)	1.3	1.4	0.09	.796	1.0	1.0	2.27	.132
ti(NgramPOSSurp,LogWordFreq)	2.8	3.5	1.09	.293	5.5	6.4	2.77	.009
ti(NgramPOSSurp,WordPos)	3.7	4.4	4.65	.001	2.1	2.7	1.85	.108
	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.718	0.053	107.87	< .001	5.772	0.050	115.69	<2e-16
SentPos	-0.004	0.007	-0.59	.558	0.019	0.005	3.47	.001
SentPos:NgramPOSSurp	-0.001	0.002	-0.27	.784	-0.005	0.002	-2.30	.022
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1395.11	< .001	41.0	41.0	2782.87	< .001
s(Trial)	54.9	62.0	21.91	< .001	61.5	62.0	555.38	< .001
s(Word)	315.7	687.0	1.31	< .001	492.7	687.0	3.31	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.979
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.95	.005
s(WordPos)	6.3	7.2	17.92	< .001	7.9	8.5	39.88	< .001
s(LogWordFreq)	1.0	1.0	0.13	.720	1.0	1.0	3.64	.056
s(WordLength)	1.0	1.0	22.27	< .001	1.0	1.0	19.29	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	0.04	.842	1.0	1.0	0.05	.833
s(NgramPOSSurp)	1.0	1.0	0.18	.672	1.0	1.0	10.94	.001
ti(NgramPOSSurp,WordLength)	1.9	2.4	1.25	.398	2.0	2.6	0.88	.537
ti(NgramPOSSurp,LogWordFreq)	1.0	1.0	0.65	.422	3.8	4.8	0.77	.551
ti(NgramPOSSurp,WordPos)	1.0	1.0	7.14	.008	1.0	1.0	1.41	.235

Table C.4: Statistical summary of the terms in the full model with n-gram word/POS surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.730	0.035	164.23	< .001	5.902	0.050	118.69	< .001
SentPos	0.008	0.006	1.33	.185	-0.047	0.007	-6.70	< .001
SentPos:NgramWordPOSSurp	0.002	0.001	2.97	.003	0.002	0.001	2.17	.030
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1077.80	< .001	32.0	32.0	1725.55	< .001
s(Trial)	61.4	62.0	231.48	< .001	61.1	62.0	208.33	< .001
s(Word)	492.5	687.0	3.54	< .001	445.9	687.0	2.57	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.936	1.0	1.0	0.45	.505
s(WordPos)	7.9	8.5	18.96	< .001	6.1	7.0	52.60	< .001
s(LogWordFreq)	1.0	1.0	4.53	.033	1.0	1.0	0.11	.745
s(WordLength)	1.0	1.0	1.99	.159	1.0	1.0	31.55	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	4.38	.036	2.8	3.2	1.99	.131
s(NgramWordPOSSurp)	3.6	4.3	1.39	.194	1.0	1.0	0.00	.979
ti(NgramWordPOSSurp,WordLength)	7.2	8.5	1.83	.053	4.1	4.4	8.36	< .001
ti(NgramWordPOSSurp,LogWordFreq)	4.2	4.8	1.71	.131	1.0	1.0	1.44	.231
ti(NgramWordPOSSurp,WordPos)	3.7	3.9	4.97	.001	1.9	2.3	0.85	.440
	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.712	0.053	107.72	< .001	5.755	0.050	115.14	< .001
SentPos	-0.029	0.007	-3.97	.000	-0.009	0.006	-1.52	.128
SentPos:NgramWordPOSSurp	0.003	0.001	3.65	.000	0.003	0.001	3.43	.001
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1395.29	< .001	41.0	41.0	2792.20	< .001
s(Trial)	55.0	62.0	21.52	< .001	61.5	62.0	584.53	< .001
s(Word)	308.3	687.0	1.24	< .001	486.3	687.0	3.21	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.979
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.95	.005
s(WordPos)	6.2	7.1	18.29	< .001	7.9	8.5	42.29	< .001
s(LogWordFreq)	1.0	1.0	0.25	.615	1.0	1.0	5.33	.021
s(WordLength)	1.0	1.0	5.11	.024	1.0	1.0	6.54	.011
ti(LogWordFreq,WordLength)	1.0	1.0	0.01	.944	1.0	1.0	0.01	.939
s(NgramWordPOSSurp)	1.0	1.0	5.37	.021	6.8	7.6	4.89	.000
ti(NgramWordPOSSurp,WordLength)	1.3	1.5	1.62	.335	1.9	2.2	0.75	.370
ti(NgramWordPOSSurp,LogWordFreq)	3.1	3.5	3.50	.009	3.8	4.7	1.61	.158
ti(NgramWordPOSSurp,WordPos)	1.1	1.2	0.02	.948	3.8	5.0	0.98	.456

Table C.5: Statistical summary of the terms in the full model with PCFG total surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.733	0.035	164.49	< .001	5.903	0.050	118.71	< .001
SentPos	0.000	0.007	-0.07	.944	-0.052	0.008	-6.54	< .001
SentPos:PCFGTotalSurp	0.004	0.001	3.91	< .001	0.003	0.001	2.52	.012
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1078.12	< .001	32.0	32.0	1723.94	< .001
s(Trial)	61.4	62.0	232.81	< .001	61.1	62.0	210.49	< .001
s(Word)	500.3	687.0	3.79	< .001	448.0	687.0	2.62	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.936	1.0	1.0	0.44	.505
s(WordPos)	8.0	8.5	16.02	< .001	6.0	6.9	50.54	< .001
s(LogWordFreq)	1.0	1.0	0.00	.971	1.0	1.0	0.01	.918
s(WordLength)	1.0	1.0	2.39	.122	1.0	1.0	32.47	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	5.72	.017	1.7	1.8	2.82	.080
s(PCFGTotalSurp)	1.0	1.0	0.36	.548	1.0	1.0	0.21	.649
ti(PCFGTotalSurp,WordLength)	2.5	2.8	1.29	.176	5.8	7.1	1.74	.089
ti(PCFGTotalSurp,LogWordFreq)	1.0	1.0	1.43	.231	1.0	1.0	0.00	.992
ti(PCFGTotalSurp,WordPos)	5.7	7.1	2.83	.006	1.0	1.0	0.31	.576
	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.712	0.053	107.69	< .001	5.763	0.050	115.39	< .001
SentPos	-0.031	0.008	-3.84	< .001	-0.014	0.006	-2.24	.025
SentPos:PCFGTotalSurp	0.004	0.001	3.45	.001	0.004	0.001	3.96	.000
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1394.68	< .001	41.0	41.0	2781.12	< .001
s(Trial)	54.9	62.0	21.42	< .001	61.5	62.0	525.41	< .001
s(Word)	307.8	687.0	1.24	< .001	489.1	687.0	3.25	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.981
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.94	.005
s(WordPos)	6.0	6.9	17.75	< .001	7.8	8.4	40.14	< .001
s(LogWordFreq)	1.0	1.0	0.31	.578	1.0	1.0	2.09	.148
s(WordLength)	1.0	1.0	2.77	.096	1.0	1.0	2.91	.088
ti(LogWordFreq,WordLength)	1.0	1.0	0.00	.994	1.0	1.0	0.43	.515
s(PCFGTotalSurp)	1.0	1.0	1.73	.188	1.0	1.0	0.60	.437
ti(PCFGTotalSurp,WordLength)	2.8	3.2	2.39	.047	1.0	1.0	3.10	.078
ti(PCFGTotalSurp,LogWordFreq)	1.2	1.3	0.75	.498	2.8	3.2	3.51	.047
ti(PCFGTotalSurp,WordPos)	1.0	1.0	1.60	.206	1.0	1.0	0.19	.664

Table C.6: Statistical summary of the terms in the full model with PCFG lexical surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.730	0.035	164.39	< .001	5.899	0.050	118.58	< .001
SentPos	0.003	0.005	0.66	.512	-0.050	0.006	-7.96	< .001
SentPos:PCFGLexSurp	0.005	0.001	4.60	< .001	0.004	0.001	3.21	.001
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1078.34	< .001	32.0	32.0	1723.74	< .001
s(Trial)	61.4	62.0	233.67	< .001	61.1	62.0	196.64	< .001
s(Word)	499.0	687.0	3.74	< .001	452.0	687.0	2.67	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.936	1.0	1.0	0.45	.505
s(WordPos)	8.0	8.5	16.88	< .001	6.1	7.0	51.80	< .001
s(LogWordFreq)	1.0	1.0	0.14	.710	1.0	1.0	1.21	.271
s(WordLength)	1.0	1.0	4.11	.043	1.0	1.0	39.46	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	3.88	.049	2.1	2.4	0.65	.552
s(PCFGLexSurp)	1.0	1.0	4.04	.044	1.0	1.0	1.57	.210
ti(PCFGLexSurp,WordLength)	3.6	4.4	1.74	.178	3.4	4.1	1.81	.122
ti(PCFGLexSurp,LogWordFreq)	1.0	1.0	0.00	.974	1.0	1.0	2.09	.149
ti(PCFGLexSurp,WordPos)	6.3	8.0	2.59	.009	1.0	1.0	0.13	.723
	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.712	0.053	107.71	< .001	5.760	0.050	115.33	< .001
SentPos	-0.024	0.007	-3.68	< .001	-0.010	0.005	-1.92	.055
SentPos:PCFGLexSurp	0.004	0.001	3.38	.001	0.004	0.001	4.41	< .001
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1392.04	< .001	41.0	41.0	2783.72	< .001
s(Trial)	54.9	62.0	21.19	< .001	61.5	62.0	541.06	< .001
s(Word)	311.3	687.0	1.30	< .001	489.2	687.0	3.26	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.980
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.95	.005
s(WordPos)	6.1	7.0	17.94	< .001	7.8	8.4	40.98	< .001
s(LogWordFreq)	1.0	1.0	0.16	.693	1.0	1.0	3.02	.082
s(WordLength)	1.0	1.0	7.94	.005	1.0	1.0	5.65	.017
ti(LogWordFreq,WordLength)	1.0	1.0	0.80	.371	1.0	1.0	0.00	.994
s(PCFGLexSurp)	1.0	1.0	3.20	.073	1.0	1.0	3.28	.070
ti(PCFGLexSurp,WordLength)	1.0	1.0	3.20	.073	1.0	1.0	2.39	.122
ti(PCFGLexSurp,LogWordFreq)	2.0	2.3	0.91	.312	3.2	3.5	3.64	.049
ti(PCFGLexSurp,WordPos)	1.0	1.0	0.73	.391	1.7	2.0	1.10	.357

Table C.7: Statistical summary of the terms in the full model with PCFG syntactic surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.743	0.035	165.11	< .001	5.909	0.050	118.88	< .001
SentPos	0.029	0.006	5.24	< .001	-0.024	0.007	-3.61	< .001
SentPos:PCFGSynSurp	-0.004	0.003	-1.47	.141	-0.006	0.003	-1.74	.082
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1082.83	< .001	32.0	32.0	1726.00	< .001
s(Trial)	61.4	62.0	232.98	< .001	61.1	62.0	196.77	< .001
s(Word)	508.7	687.0	3.94	< .001	451.7	687.0	2.69	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.935	1.0	1.0	0.45	.505
s(WordPos)	7.9	8.5	16.73	< .001	6.1	7.0	53.27	< .001
s(LogWordFreq)	1.0	1.0	1.74	.187	1.0	1.0	2.70	.100
s(WordLength)	1.0	1.0	12.20	< .001	1.0	1.0	43.78	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	0.29	.591	3.3	3.6	1.62	.173
s(PCFGSynSurp)	1.0	1.0	5.97	.015	1.0	1.0	9.66	.002
ti(PCFGSynSurp,WordLength)	1.8	2.1	1.06	.389	1.0	1.0	0.07	.788
ti(PCFGSynSurp,LogWordFreq)	1.0	1.0	2.85	.091	1.0	1.0	0.42	.519
ti(PCFGSynSurp,WordPos)	3.1	3.5	6.46	< .001	1.7	2.1	3.21	.035

	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.718	0.053	107.88	< .001	5.772	0.050	115.67	< .001
SentPos	-0.005	0.007	-0.82	.413	0.014	0.005	2.65	.008
SentPos:PCFGSynSurp	0.000	0.003	-0.09	.932	-0.004	0.003	-1.30	.193
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1395.06	< .001	41.0	41.0	2789.86	< .001
s(Trial)	54.9	62.0	21.71	< .001	61.5	62.0	571.90	< .001
s(Word)	311.3	687.0	1.27	< .001	492.4	687.0	3.32	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.979
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.94	.005
s(WordPos)	6.1	7.0	18.06	< .001	7.8	8.4	39.16	< .001
s(LogWordFreq)	1.0	1.0	0.18	.674	1.0	1.0	3.75	.053
s(WordLength)	1.0	1.0	21.70	< .001	1.0	1.0	18.48	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	0.00	.996	1.0	1.0	0.07	.792
s(PCFGSynSurp)	1.0	1.0	1.43	.233	1.0	1.0	7.50	.006
ti(PCFGSynSurp,WordLength)	2.3	2.7	1.42	.168	1.0	1.0	0.23	.630
ti(PCFGSynSurp,LogWordFreq)	2.3	2.6	2.76	.039	6.5	7.6	2.81	.004
ti(PCFGSynSurp,WordPos)	2.5	2.9	2.54	.038	6.7	8.2	4.23	< .001

Table C.8: Statistical summary of the terms in the full model with PCFG POS surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.742	0.035	165.04	< .001	5.906	0.050	118.77	< .001
SentPos	0.025	0.006	4.42	< .001	-0.030	0.007	-4.34	< .001
SentPos:PCFG_POSSurp	-0.001	0.002	-0.54	.591	-0.001	0.002	-0.56	.573
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1078.43	< .001	32.0	32.0	1718.02	< .001
s(Trial)	61.4	62.0	229.88	< .001	61.1	62.0	208.74	< .001
s(Word)	503.2	687.0	3.83	< .001	449.1	687.0	2.62	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.936	1.0	1.0	0.45	.505
s(WordPos)	8.0	8.5	15.88	< .001	6.1	7.0	43.06	< .001
s(LogWordFreq)	1.0	1.0	2.75	.097	1.0	1.0	3.67	.055
s(WordLength)	1.0	1.0	14.23	< .001	1.0	1.0	39.11	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	0.14	.706	3.6	4.0	2.44	.054
s(PCFG_POSSurp)	1.0	1.0	0.34	.563	3.1	3.7	1.83	.107
ti(PCFG_POSSurp,WordLength)	8.4	9.5	2.75	.003	1.0	1.0	0.02	.882
ti(PCFG_POSSurp,LogWordFreq)	9.4	10.6	3.93	< .001	8.4	9.6	2.90	.002
ti(PCFG_POSSurp,WordPos)	3.4	3.7	3.84	.051	4.7	6.0	1.67	.142
	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.719	0.053	107.90	< .001	5.773	0.050	115.68	< .001
SentPos	-0.007	0.007	-0.99	.324	0.018	0.006	3.30	.001
SentPos:PCFG_POSSurp	0.000	0.002	0.17	.865	-0.004	0.002	-2.08	.038
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1394.57	< .001	41.0	41.0	2770.81	< .001
s(Trial)	54.8	62.0	21.88	< .001	61.5	62.0	531.48	< .001
s(Word)	310.7	687.0	1.26	< .001	487.5	687.0	3.20	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.980
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.94	.005
s(WordPos)	6.2	7.1	17.34	< .001	7.9	8.5	35.39	< .001
s(LogWordFreq)	1.0	1.0	0.25	.619	1.0	1.0	3.99	.046
s(WordLength)	1.0	1.0	24.27	< .001	1.0	1.0	20.00	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	0.20	.654	1.0	1.0	0.03	.860
s(PCFG_POSSurp)	1.6	2.0	1.14	.296	1.0	1.0	4.68	.030
ti(PCFG_POSSurp,WordLength)	1.0	1.0	2.65	.103	4.0	5.0	1.57	.172
ti(PCFG_POSSurp,LogWordFreq)	6.6	8.0	1.90	.050	8.1	9.4	2.54	.006
ti(PCFG_POSSurp,WordPos)	1.0	1.0	4.71	.030	9.5	10.8	2.19	.008

Table C.9: Statistical summary of the terms in the full model with RNNG surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.736	0.035	164.45	< .001	5.904	0.050	118.73	< .001
SentPos	0.009	0.005	1.66	.097	-0.042	0.007	-6.33	< .001
SentPos:RNNGSurp	0.002	0.001	3.00	.003	0.002	0.001	1.61	.108
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1078.24	< .001	32.0	32.0	1726.24	< .001
s(Trial)	61.4	62.0	232.51	< .001	61.1	62.0	196.48	< .001
s(Word)	500.3	687.0	3.75	< .001	448.3	687.0	2.60	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.935	1.0	1.0	0.45	.505
s(WordPos)	7.9	8.5	17.19	< .001	6.1	6.9	49.49	< .001
s(LogWordFreq)	1.0	1.0	0.03	.868	1.0	1.0	0.06	.802
s(WordLength)	1.0	1.0	5.86	.015	1.0	1.0	38.85	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	3.46	.063	1.0	1.0	9.54	.002
s(RNNGSurp)	1.0	1.0	0.70	.402	1.0	1.0	0.20	.659
ti(RNNGSurp,WordLength)	2.6	3.2	1.31	.252	5.5	6.5	3.73	.001
ti(RNNGSurp,LogWordFreq)	3.7	4.4	2.11	.069	1.7	2.0	0.10	.888
ti(RNNGSurp,WordPos)	4.5	5.5	2.95	.006	1.0	1.0	0.37	.546
	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.714	0.053	107.80	< .001	5.765	0.050	115.42	< .001
SentPos	-0.022	0.007	-3.20	.001	-0.007	0.005	-1.39	.166
SentPos:RNNGSurp	0.003	0.001	2.87	.004	0.003	0.001	3.61	< .001
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1394.53	< .001	41.0	41.0	2794.45	< .001
s(Trial)	54.8	62.0	21.97	< .001	61.5	62.0	594.77	< .001
s(Word)	310.7	687.0	1.29	< .001	487.8	687.0	3.23	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.979
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.95	.005
s(WordPos)	6.1	7.0	16.96	< .001	7.9	8.4	38.97	< .001
s(LogWordFreq)	1.0	1.0	1.45	.228	1.0	1.0	0.74	.388
s(WordLength)	1.0	1.0	7.47	.006	1.0	1.0	7.11	.008
ti(LogWordFreq,WordLength)	1.0	1.0	0.02	.902	1.0	1.0	1.27	.260
s(RNNGSurp)	1.0	1.0	0.07	.790	1.0	1.0	1.94	.164
ti(RNNGSurp,WordLength)	1.0	1.0	0.73	.392	2.3	2.8	0.37	.693
ti(RNNGSurp,LogWordFreq)	1.0	1.0	4.79	.029	3.2	3.8	2.44	.107
ti(RNNGSurp,WordPos)	1.0	1.0	3.66	.056	1.0	1.0	0.79	.375

Table C.10: Statistical summary of the terms in the full model with transformer surprisal.

	English				**Chinese**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.738	0.035	164.54	< .001	5.903	0.050	118.70	< .001
SentPos	0.008	0.006	1.48	.138	-0.046	0.007	-6.87	< .001
SentPos:TransformerSurp	0.003	0.001	3.06	.002	0.002	0.001	2.33	.020
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	30.0	30.0	1076.87	< .001	32.0	32.0	1724.18	< .001
s(Trial)	61.4	62.0	229.52	< .001	61.1	62.0	207.68	< .001
s(Word)	501.4	687.0	3.73	< .001	447.5	687.0	2.59	< .001
s(VocabAdjPerf)	1.0	1.0	6.69	.010	1.0	1.0	0.04	.852
s(ReadingCompAdjPerf)	1.0	1.0	0.01	.936	1.0	1.0	0.45	.505
s(WordPos)	7.9	8.5	16.85	< .001	6.0	6.9	49.64	< .001
s(LogWordFreq)	1.0	1.0	0.13	.718	1.0	1.0	0.10	.749
s(WordLength)	1.0	1.0	5.69	.017	1.0	1.0	47.46	< .001
ti(LogWordFreq,WordLength)	1.0	1.0	1.89	.169	1.0	1.0	3.62	.057
s(TransformerSurp)	1.0	1.0	0.75	.388	1.0	1.0	0.00	.957
ti(TransformerSurp,WordLength)	1.0	1.0	0.39	.532	6.4	7.8	3.03	.003
ti(TransformerSurp,LogWordFreq)	3.7	4.6	2.43	.050	2.2	2.4	2.31	.248
ti(TransformerSurp,WordPos)	6.0	7.4	3.04	.003	1.0	1.0	0.00	.987

	Korean				**Spanish**			
	Estimate	***SE***	***t***	***p***	**Estimate**	***SE***	***t***	***p***
Intercept	5.714	0.053	107.76	< .001	5.766	0.050	115.49	< .001
SentPos	-0.024	0.007	-3.44	.001	-0.004	0.005	-0.74	.461
SentPos:TransformerSurp	0.003	0.001	3.12	.002	0.002	0.001	2.87	.004
	Eff. df	**Ref. df**	***F***	***p***	**Eff. df**	**Ref. df**	***F***	***p***
s(Participant)	25.0	25.0	1392.99	< .001	41.0	41.0	2795.95	< .001
s(Trial)	54.8	62.0	20.78	< .001	61.5	62.0	607.40	< .001
s(Word)	307.5	687.0	1.22	< .001	490.3	687.0	3.30	< .001
s(VocabAdjPerf)	1.0	1.0	0.00	.963	1.0	1.0	0.00	.978
s(ReadingCompAdjPerf)	1.0	1.0	0.20	.655	1.0	1.0	7.95	.005
s(WordPos)	6.1	7.0	17.14	< .001	7.9	8.5	40.54	< .001
s(LogWordFreq)	1.0	1.0	1.03	.310	1.0	1.0	1.41	.235
s(WordLength)	1.0	1.0	4.69	.030	1.0	1.0	7.88	.005
ti(LogWordFreq,WordLength)	1.0	1.0	0.01	.926	1.0	1.0	1.56	.211
s(TransformerSurp)	1.0	1.0	0.93	.335	1.0	1.0	0.36	.551
ti(TransformerSurp,WordLength)	4.3	5.6	1.06	.343	1.0	1.0	0.26	.609
ti(TransformerSurp,LogWordFreq)	1.0	1.0	0.37	.546	1.5	1.7	1.82	.251
ti(TransformerSurp,WordPos)	1.0	1.0	1.50	.221	2.7	3.1	2.30	.154