# Multi-Word Representations in Minds and Models:
# Investigating the Storage of Multi-Word Phrases in Humans and Large Language Models

By

Zachary Nicholas Houghton

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Linguistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

______________________________________

Dr. Emily Morgan, Chair

______________________________________

Dr. Masoud Jasbi

______________________________________

Dr. Fernanda Ferreira

Committee in Charge

2024

# Table of contents

# List of Figures

# List of Tables

# Abstract

This is my abstract.

# Acknowledgements

First and foremost, I would not be here if it weren't for my advisor, Dr. Emily Morgan. Emily has been a never-ending source of knowledge and a constant source of reassurance. Emily was charged with the non-trivial task of helping to translate my incoherent stream of thoughts into a coherent set of ideas. She pushed me hard, believed in me, and never let me fall behind. I'm extraordinarily fortunate to have had her as my advisor.

I'd also like to thank many of the other professors here who have been crucial to my development as a linguist. Specifically, I'd like to thank Dr. Fernanda Ferreira whose knowledge of the field is so expansive that she can, without delay, provide a reference for any psycholinguistic phenomenon one can think of. I'd also like Dr. Kenji Sagae for his help over the years regarding all things natural language processing. Additionally, I'd like to thank Dr. Santiago Barreda who has provided a great deal of advice and help with statistical analyses. Finally, I'd like to thank Dr. Masoud Jasbi who has taught me the importance of various linguistics theories, even those I don't agree with.

Many of the ideas presented here have benefited in some form or another from feedback from many brilliant graduate students. I would especially like to thank Dingyi (Penny) Pan and Casey Felton for their feedback on much of the work included here.

I'd also like to thank Casey, Felix, and Nora for being a strong support system during my time here. Our Sunday shenanigans were a welcomed escape from the tireless work of completing a PhD.

My journey in linguistics started at the University of Oregon, and I want to thank all of the professors that supported the beginning of my journey. I particularly want to acknowledge Dr. Vsevolod Kapatsinski. Volya has donated countless hours of his time to me even after his role as my undergraduate thesis advisor was long over. He continues to be an endless source of knowledge and inspiration and much of my knowledge and interest in language learning comes from him. Perhaps more importantly, however, he is a constant reminder that linguistics is *fun*! Had it not been for our meetings over the years that devolved into ridiculous linguistic tangents, I would have burnt out long ago.

I would also like to thank Kim 선생님 who encouraged me to apply for graduate school in the first place and believed in me often times more than I believed in myself.

In addition, I want to thank Dr. Melissa Baese-Berk and Dr. Misaki Kato, and Dr. Zara Harmon. A great deal of my success as a graduate student came from their mentorship.

Along with the technical and academic guidance, it also would have been impossible to complete this PhD without the unending support I received from so many people in my personal life.

Specifically, I have been fortunate to have a strong support system in the form of of my two sisters, Kayla and Lily. We've been through so much together. I don't know where I would be, not just academically, but more generally in life, had you two not been by my side.

This work would also have not been completed without the influence of my parents. Specifically, I want to thank my mom for teaching me that the ability to find the answer is far more important than knowing the answer, and my dad, for teaching me the discipline and practical skills to achieve my goals.

For both my undergraduate and graduate studies, I made the difficult decision to pursue my education on the opposite side of the country from my hometown in Connecticut. Despite the physical distance, I have always been able to count on my friends back east—not only during the easier times, but, more importantly, through the more challenging ones. Addy, Charles, Paul, Spencer, Wyatt, and 보미: thank you for being such a strong and constant support network.

Finally, to all the people I met during these five years at UC Davis: you welcomed me and gave me a home. You supported me, believed in me, and pushed me to be the best version of myself. I'm exceptionally proud to be an Aggie.

The number of people who have been indispensable in me getting here is undoubtedly larger than is feasible to include here. To those that I have inevitably left out, I apologize.

# Chapter 1

# Introduction

How much of language is memorized and how much is improvised? Every time we speak, we are faced with the choice between familiar expressions, like *I don't know*, and novel constructions, like *to me it is uncertain*. In other words, we constantly navigate a trade-off between stored, item-specific knowledge – our stored knowledge of particular words and phrases – and abstract knowledge, which allows us to combine those stored representations in new ways.

From a young age, humans are capable of generating sentences that we've never encountered before (Berko, 1958). This ability is largely enabled by an ability to store forms (storage) that we've learned and combine them (computation) into new forms using our knowledge of the grammar (Berko, 1958; Morgan & Levy, 2015a, 2016a, 2024; Stemberger & MacWhinney, 1986, 2004). In theory, storage and computation can be complementary forces: if a form is stored, it does not necessarily need to be computed, and if a meaning can be generated via computation, it does not necessarily need to be stored. For example, if the word *cats* is stored, then it is not necessary to compute it (e.g., by combining the lexical root, *cat*, with a general plural rule, *-s*). On the other hand, if it can be computed (e.g., if we have learned the word *cat* and we have learned how to make regular forms plural in English), then we do not necessarily need to store it. However, the fact that computation and storage can be independent does not preclude the possibility of items being both stored holistically and able to be formed compositionally. Indeed, a surge of research in the last few decades has suggested the opposite: that a rich amount of language, including words and multi-word phrases, is stored holistically (e.g., Bybee, 2003; Morgan & Levy, 2015b, 2016a, 2024; Stemberger & MacWhinney, 1986, 2004).

The evidence that more complex forms, such as multi-morphemic words and phrases, may be stored holistically begs a host of questions. For example, what factors determine whether a phrase is stored holistically or generated compositionally using knowledge of the grammar? If *pick up* is stored holistically, then under what circumstances does a listener use their knowledge of the grammar to form the phrase compositionally and under what circumstances do they access the holistically stored representation? Similarly, how do compositional forms interact with their holistically stored counterparts during processing? When listeners hear *pick*, are they able to activate the representation of the holistically stored form *pick up* before hearing *up*? Finally, how do stored representations differ from the individual word-level representations? Is the representation of *pick up* completely disconnected from the individual representations of *pick* and *up*, despite the fact that they have overlapping acoustics (i.e., when hearing *pick up* one necessarily hears *pick* and *up*)?

The present section introduces the relevant background for each of these questions. Section 1.1 describes the current debates about storage. Section 1.2 explores the evidence for the holistic storage of multi-morphemic words and multi-word phrases. Section 1.3 reviews the evidence for factors that drive storage. Section 1.4 examines how stored items are represented. Section 1.5 examines the processing consequences of holistic storage. Section 1.6 examines how large language models trade off between storage and computation. Finally, Section 1.7 outlines the rest of the dissertation.

## 1.1 Accounts of Storage

Traditional linguistic theories have assumed that very little is stored and instead a great deal of language production leverages humans' remarkable ability to generate complex meaning by composing words together (Chomsky, 1965). This was based on an assumption that human memory is limited and storing items that could be generated compositionally would be an inefficient use of memory. These theories posit that stems of words are stored and more complex word forms are generated via regular rules. For example, *cat* would be stored and *cats* would be generated using knowledge of the grammar. Similarly, phrases would be generated so long as they're compositional. Holistic storage of multi-word phrases is instead reserved for idioms (Chomsky, 1965) and perhaps extremely high-frequency items (Pinker & Ullman, 2002). According to these theories, *I don't know* would be generated by accessing the individual words and then combining the individual representations together.

However, there may be advantages to storing words or phrases that we can compute. For example, if we are producing a combination of words often enough (e.g., *bread and butter*), it may be efficient to store it in memory and retrieve the stored representation instead of composing it every time. Further, the brain may have dramatically more space for storage than we had previously realized, with an upper bound of $10^{8432}$ bits (Wang et al., 2003). Given that Mollica & Piantadosi (2019) have estimated that in terms of linguistic information humans store only somewhere between one million and ten million bits of information, memory constraints are not likely to be a limitation to this.

Following this, usage-based theories have long posited the possibility of multi-word phrases being stored holistically (Bybee, 2003; e.g., Bybee & Hopper, 2001; Kapatsinski, 2018; Morgan & Levy, 2015b, 2016a, 2024; Stemberger & MacWhinney, 1986, 2004). These theories posit that multi-word phrases can be stored if they're used often enough. For example Tomasello (2005) argued that early verb knowledge is holistic in nature, with children reproducing memorized chunks as opposed to generating verbs in novel contexts. Further, Bybee (2003) argued that after learning to produce these verbs in novel contexts, children don't necessarily flush these holistic representations from memory. Instead, proponents of usage-based theories argue that high-frequency phrases like *I don't know* are stored holistically while lower and mid-frequency phrases are generated compositionally.

Usage-based theories of storage naturally developed out of the phonetics and phonology literature, being championed by linguists such as Janet Pierrehumbert and Joan Bybee, who demonstrated that phonetic representations could not be reduced to abstract representations with no phonetic details (Bybee, 2002, 2003; Bybee & Scheibman, 1999; e.g., Pierrehumbert, 2016). Instead, abstract representations require some link to phonetic details in various contexts. In other words, the pronunciation for a word cannot be simply reduced to individual phonemes, because the pronunciations for those phonemes depend on various factors, such as the frequency of the word and the adjacent phonemes. For example, Bybee (2002) demonstrated that phonetically conditioned changes affect high-frequency words before low-frequency words. She further demonstrated that a word that occurs more often in a context favorable for the phonetically conditioned change will change more quickly. She argued that it is difficult to account for these with a model that reduces words to abstract phonological representations that are context-independent. Similarly, McMurray et al. (2009) demonstrated that people are sensitive to gradient changes in VOT. However if people are representing /b/ and /p/ as abstract phonetic categories, than people should not be sensitive to within-category variability.

The phonetics literature demonstrated that representations of words could not be reduced to context-independent phoneme-level representations and this lead to many of the same questions being asked about higher levels of representations (e.g., Bybee, 2003). That is, if simple words are being represented holistically with rich phonetic detail, then perhaps it is possible that similarly monomorphemic words or even phrases may also be represented holistically.

## 1.2 Evidence of Storage in Words and Phrases

There is a great deal of evidence that multi-word phrases are stored holistically. For example, high-frequency phrases such as *I don't know* undergo phonetic reduction that isn't seen in other low or mid-frequency phrases containing *don't* (Bybee & Scheibman, 1999). If the representation of *don't* is the same across different context, we would expect *don't* to be equally reduced in those contexts. As such, the phonetic reduction of *I don't know* suggests a holistic representation separate from that of the individual words. Following this, Bybee (2003) demonstrated that there are many high-frequency phrases that undergo phonetic reduction that can't be accounted for by phonetic reduction of the words outside of those phrases ( e.g., *going to, have to, want to*, etc).

Similarly, Yi (2002) found evidence for holistic storage of phrases in Korean as well. In Korean, certain consonants undergo tensification when they occur after the future tense marker *-l*. Yi (2002) demonstrated that the rate of this tensification is higher in high-frequency phrases than low-frequency phrases, suggesting that they have a separate representation. These results parallel findings on a word-level (which most theories posit have separate representations). For example, in Korean epenthesis (insertion of a sound) occurs more often in high-frequency words than in low-frequency words (Lee & Kapatsinski, 2015). Similarly, deletion is more likely to occur in a frequent word like *most* than in an infrequent word like *mast* (Bybee, 2002; Kapatsinski, 2021). This parallelism is important because monomorphemic words must be stored. Thus the fact that the patterns of phonetic reduction in certain phrases mirrors the patterns at the word-level suggest that they may be stored.

The psycholinguistics literature has also provided an abundance of evidence for multi-word holistic storage Stemberger & MacWhinney (1986). For example, by examining corpus data Stemberger & MacWhinney (1986) demonstrated that errors occur less in high-frequency words than low-frequency words. They argued that one of the consequences of high-frequency is greater accuracy. They further suggested that if inflected forms are stored holistically than no-marking errors should be less common in high-frequency inflected forms than in lower frequency forms for both regular and irregular items. This is exactly what they found: they showed that fewer no-marking errors (e.g., producing *walk* instead of *walked*) are made on the past-tense forms of frequent verbs relative to infrequent verbs. However, corpus data can be messy, so they ended their study by examining spontaneous speech. They found that in spontaneous speech participants produce no-marking errors less often in high-frequency regular verbs than in low-frequency regular verbs. They argued that if people are accessing each morpheme (e.g., accessing *walked* by accessing *walk*, then *-ed*), then errors on *-ed* should be independent of the base form. That is, accessing *walk* more easily or more difficulty should not influence the error rate of the past-tense morpheme, which is constant across all verbs (if they're not stored holistically).

In addition to production, the processing literature has also found evidence of storage. For example, Siyanova-Chanturia et al. (2011) investigated the reading times of binomials in their frequent ordering (e.g., *bread and butter*) and in their infrequent ordering (e.g., *butter and bread*). They found that humans read the binomial faster in their frequent ordering. Further, in a follow-up study Morgan & Levy (2015b) examined whether this finding is due to generative constraints, such as a preference for short words before long words (Benor & Levy, 2006) or whether it is due to the items being stored holistically. By annotating a corpus of binomials for constraints known for affecting the orderings of binomials in corpus data, Morgan & Levy (2016a) developed a probabilistic model to predict binomial orderings. The model combines various constraints that affect binomial orderings into a single preference estimate that indicates the preferred order and the strength of that preference for a given binomial (i.e., whether *bread and butter* is preferred over *butter and bread*, and by how much). Further, Morgan & Levy (2016a) demonstrated that this generative preference value is a strong predictor of human ordering preferences for lower-frequency items, but not for high-frequency items, suggesting that humans rely primarily on item-specific knowledge for high-frequency items. Interestingly though, in a follow-up study Morgan & Levy (2024) demonstrated that these generative preferences exert an effect on all items, regardless of frequency, although they are weaker for high-frequency items.

A related line of research examined the role of storage in the development of frequency-dependent preference extremity, which is the tendency of high frequency items to be more polarized in their orderings than low frequency items (Liu & Morgan, 2020, 2021; Morgan & Levy, 2016b). For example, Morgan & Levy (2016b) demonstrated that high frequency binomials (e.g., *bread and butter*) are more fixed in their ordering preferences than low or mid frequency binomials. Similar work has also demonstrated that more frequent verbs have more polarized preferences with respect to the dative alternation (Liu & Morgan, 2020) and that adjectives in adjective-adjective-noun constructions (big round ball) show more polarized preferences (Liu & Morgan, 2021). Further, Morgan & Levy (2016b) demonstrated that a model that assumes binomials are stored holistically predicts the emergence of these preferences over generations of learners. It is harder to account for this pattern without storage at the phrase level, because generative preferences do not predict the ordering preferences for high-frequency items. Thus, if people are simply composing binomials on the fly from their individual words, it's unclear why high-frequency items become polarized in their orderings.

Finally, there is also evidence of holistic storage from the learning literature O'Donnell (2016). For example, there is evidence that attending to the whole utterance as opposed to attending to each individual word facilitates learning (Siegelman & Arnon, 2015). Specifically, Siegelman & Arnon (2015) gave adult L2 learners of German sentences that were either segmented into individual words, or not segmented at all. They found that participants learned grammatical gender in German better when the sentences were not segmented. They argued that by presenting participants with the unsegmented segments, participants were forced to pay attention to bigger chunks of the sentences, making it easier for them to learn the grammatical gender patterns. This suggests that holistic storage may actually facilitate the learning of various grammatical relationships.

Additionally, in modeling learning of the English past tense, models that store some items holistically outperform models that don't (O'Donnell, 2016). For example, O'Donnell (2016) tested 4 probabilistic models on their ability to learn the English past tense. These models differed in whether they stored items holistically or composed them using morphological knowledge. He found that their inference-based model, which stored units of varying sizes, was able to learn the English past tense much better than the other models.

However, while there is an abundance of evidence that a lot more is stored than was previously considered, what factors determine whether a multi-word phrase is stored holistically?

## 1.3 Factors that Drive Storage

Despite the evidence of holistic storage, it is still unclear what drives storage. Frequency has been assumed to be the driving force of storage in the literature, and for good reason: there is no shortage of evidence that frequency drives holistic storage (Bybee, 2003; Bybee & Hopper, 2001; Bybee & Scheibman, 1999; Kapatsinski & Radicke, 2009; Morgan & Levy, 2015b, 2016a; Pierrehumbert, 2016; Stemberger & MacWhinney, 1986, 2004). For example, as mentioned in the previous section phonetic reduction has been shown to occur in high frequency multi-word phrases, but not in their medium/low frequency counterparts (Bybee, 2003; Bybee & Scheibman, 1999). In addition to previous examples, there is also evidence that high frequency phrases lose the recognizability of their component parts relative to low or mid frequency phrases (Kapatsinski & Radicke, 2009). In other words, *up* is harder to recognize in *pick up* than in *run up* (we will revisit this example in more detail in the next section).

While frequency clearly plays a large role in holistic storage, it's unclear if other factors also influence whether a word or phrase is stored holistically. For example, predictability also plays an important part in many linguistic theories, especially in the learning literature (Kapatsinski, 2018; Olejarczuk et al., 2018; Ramscar et al., 2013; Saffran et al., 1996). Olejarczuk et al. (2018) examined how humans learn new phonetic categories. They found that when learners experience a rare member of a phonetic category, that member of the category exerts a disproportionate influence on the learner's representation of that category. In other words, learners try to actively predict upcoming sounds when learning new phonetic categories and update their representations in proportion to how surprising the upcoming sound is (Olejarczuk et al., 2018).

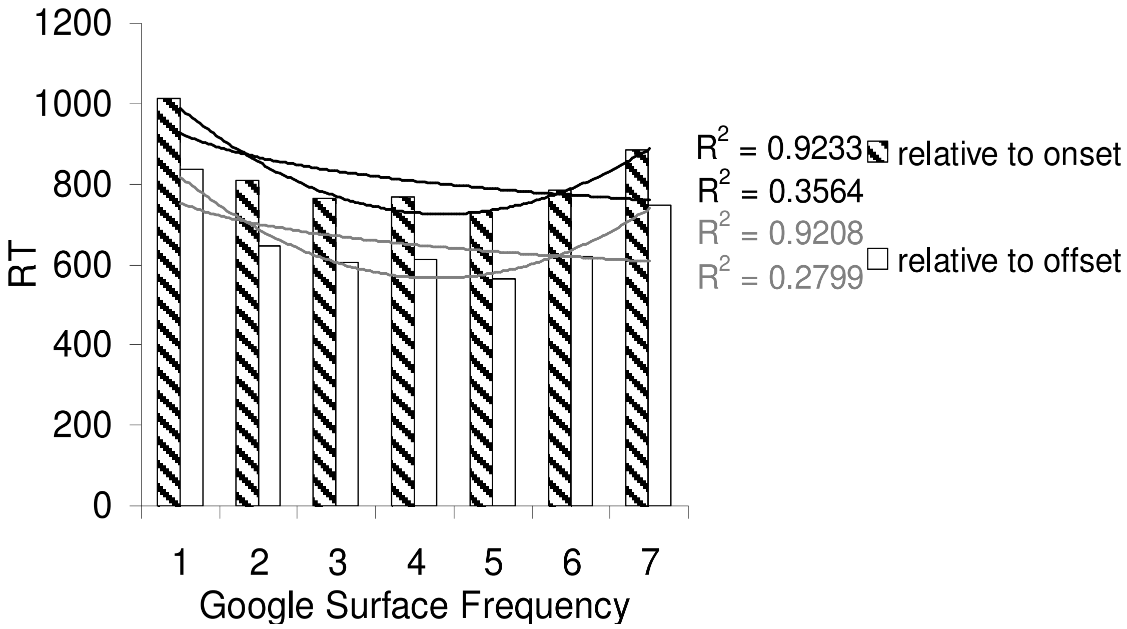
Additionally, Ramscar et al. (2013) demonstrated that children rely on how predictive a word is of a meaning when learning the meanings of words. Specifically, they examined how children and adults learn novel words. In their artificial language paradigm, participants first saw two objects together (A and B) and heard them labeled ambiguously as a *dax*. They then heard B and C together with the ambiguous label, *pid*. In testing, all three objects were presented, and participants were tasked with identifying the *dax*, the *pid*, and the *wug*, which was a novel label. They found that both children and adults learn that A refers to *dax* and C refers to *pid*, but differ in what they learn *wug* refers to. Adults use logical exclusion and learn that *wug* refers to *B*, however children learn that B is not as predictive of *dax* or *wug*, but do not conclude that B must then refer to *wug*. Instead, since B has a higher background rate (it occurs in every context), children learn that *wug* refers to either A or C.

Finally, learners are highly sensitive to predictability when segmenting the speech stream into words (Saffran et al., 1996). In their classic paper, Saffran et al. (1996) demonstrated that children leverage transitional probabilities to segment the speech stream into individual words. These results, taken in conjunction with the other results demonstrating the importance of predictability in learning, suggest that predictability may also drive what we store holistically.

## 1.4 Representations of Stored Items

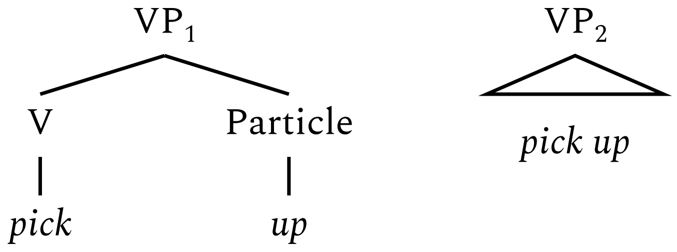
An equally important question is how are stored items represented. Specifically, do stored representations maintain internal structure with respect to their component parts? For example, Kapatsinski & Radicke (2009) argued that stored constructions may lose some amount of their internal structure. They presented participants with sentences containing *up* either inside of a word (e.g., *cup*) or inside of a V+*up* construction (e.g., *pick up*). Participants were tasked with pressing a button when they heard *up*. They found that in high frequency V+*up* constructions it is harder to recognize *up* than in medium frequency phrases, even after accounting for phonetic reduction (Figure 3.1.3.1). In other words, participants grow faster to recognize up as the frequency of the phrase increases, until reaching the highest frequency phrases, where their reaction time grows slower. This increase in recognition time suggests 1) that high frequency V+*up* phrases are stored holistically (since participants should be faster to recognize words in high-frequency contexts if they are forming them compositionally) and 2) holistically stored items lose some amount of their internal structure.

Figure 1.4.0.1: A plot of the results reproduced from Kapatsinski & Radicke (2009).



A visualization of what this may look like is demonstrated in Figure 1.4.0.2. The left tree represents the phrase *pick up* stored holistically but with intact internal structure and the right tree represents the phrase *pick up* stored holistically but without internal structure.

Figure 1.4.0.2: A visualization of a holistically stored phrase with internal structure (left) and without internal structure (right).



This lack of internal structure could be lost over time, or it may simply not have been learned in the first place. A great deal of children's early learning is driven by memorizing chunks (Bybee, 2003; Tomasello, 2005). For example, Tomasello (2005) argued that young children learn verbs in fixed "islands", producing them in fixed-constructions before eventually learning to generalize them to other contexts. As a result, children may be learning holistic representations of them initially.

Additionally, if predictability drives word-segmentation, many predictable phrases may be segmented out of the speech stream as a single chunk. Following this, Bybee (2003) argued that after learning these chunks, it seems unlikely that children would then flush these from their memory. Further, many high frequency and high predictability phrases have semantically vague relationships (e.g., *trick or treat*). These phrases may be difficult to breakdown into their component words due to the lack of semantic transparency. Children may learn to store these phrases holistically. However, it is possible that their representations for these items are updated to reflect knowledge of the individual words upon further learning. Thus it is not entirely clear if holistically stored items lack internal representations of the individual words.

On the other hand, internal structure could be lost over time. For example, learners are more likely to semantically extend frequent forms to novel contexts than infrequent forms (Harmon & Kapatsinski, 2017). This increase in accessibility may similarly drive their loss of internal structure over time: as a phrase is extended to new contexts, the representation of that phrase may also become more general to accommodate the new context. This may lead to the internal structure being lost over time as the contexts that the phrase is used in becomes more different from the contexts in which the individual words are used.

## 1.5 Processing Consequences of Storage

Speech is inherently temporally linear: unlike reading, when you hear a sentence, you cannot skip forward or rewind back in time. As such, how are holistically stored multi-word phrases processed? One possibility is that when hearing part of the phrase, listeners may access the representation for the holistically stored phrase. For example, in an extreme case, hearing *Habeas* may be enough to access the representation of *Habeas Corpus*.

However, this seems not to be the case. For example, Staub et al. (2007) examined the effects of plausibility on the reading times of high frequency and low frequency compound nouns (Noun and Noun compounds). Specifically, participants read sentences that were either locally plausible or locally implausible:

1. **Novel Compound**

   **1a** The zookeeper picked up the monkey medicine that was in the enclosure.

   **1b** The zookeeper spread out the monkey medicine that was in the enclosure.

2. **Familiar Compound**

   **2a** Jenny looked out on the huge mountain lion pacing in its cage.

   **2b** Jenny heard the huge mountain lion pacing in its cage.

For example, Sentence 1a is locally plausible because the sentence is plausible at the first noun. That is, *picked up the monkey* is plausible. On the other hand, 1b is locally implausible because it is implausible at the first noun; *spread out the monkey* is not plausible. Sentences 2a and 2b are analogous but with a high frequency compound noun.

Staub et al. (2007) examined readers' eye-movements as they read these sentences and found that for locally implausible sentences there was a slowdown at the first noun. Crucially, this slowdown was equal for both the high and low frequency compound nouns. However, if high frequency compound nouns are stored holistically, and humans are able to access the representation at the first noun, then participants should have been able to overcome at least some of the slowdown due to the implausibility effect for the high frequency items. However, this is not what Staub et al. (2007) found.

Given the results of Staub et al. (2007), one natural possibility is that recognition happens incrementally and the representation of the phrase becomes activated more strongly than the words after the listener has heard each of the words in the phrase. However, if this was the case then the results from Kapatsinski & Radicke (2009) discussed earlier would complicate things. Recall that Kapatsinski & Radicke (2009) found that participants are slower to recognize *up* in high frequency phrases. If recognition occurs incrementally, one would expect that for high frequency phrases, *up* would be recognized even faster. That is, a slower recognition of *up* in the context of high-frequency phrases indicates that *up* is harder to recognize depending on what the preceding verb is. It's hard to see how an incremental approach can account for this context-dependent difficulty in recognition.

Following these results, it is possible that an incremental approach with competition (such as one proposed by McClelland et al., 1984) may be the best account. Specifically, it is possible that processing unfolds incrementally such that the representations of each individual word are activated, however upon accessing the representation of the phrase the word-level representations may be inhibited. For example, upon hearing *pick* perhaps the word-level representation for *pick* is activated, and upon hearing *up* perhaps instead of activating the representation for *up*, the holistically stored representation for *pick up* is activated and the representations at the word-level (*pick* and *up*) are inhibited. This inhibition of *up* maybe be the cause of the increased recognition times for high-frequency V+*up* compounds.

## 1.6 Storage in Humans vs Large Language Models

By now it should be clear that the literature has demonstrated that a great deal of items are stored holistically. However it's unclear how this can arise as a function of learning. In order to help address this question, we turn to large language models, which have developed rapidly over the last several years.

### 1.6.1 Transformer Model Architecture

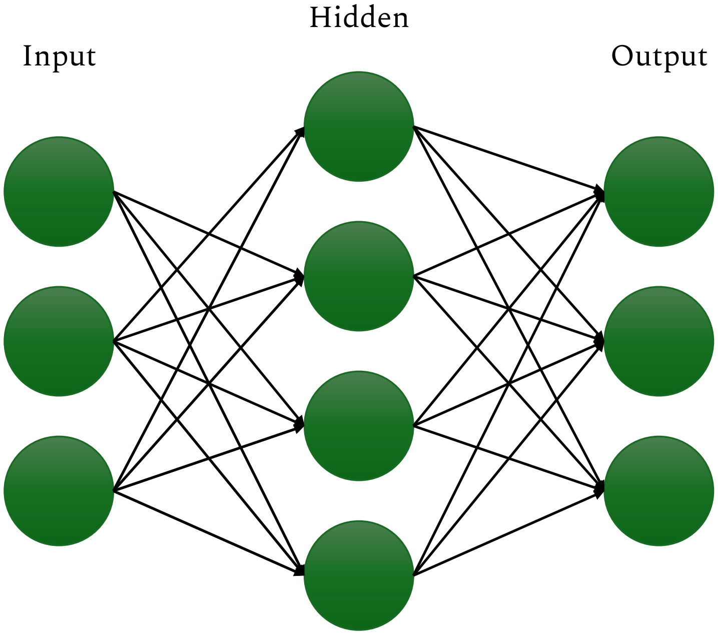
When I started this program in 2020, the idea of a single language model that could produce fluent text that was discernible from text written by humans seemed like a far-off dream. However, with rapid advancements in transformer models, the language models of today have seemed to accomplish that goal. With their rapid advancement, the question of whether they may be accomplishing this goal in a similar manner as humans has been at the forefront of a great deal of linguistics and cognitive science research. Thus in this section, we will introduce the transformer architecture along with the current state of the literature on their ability to store and learn generalized knowledge.

The term large language model typically refers to a transformer model (Vaswani et al., 2017), such as Llama (Touvron et al., 2023). The heart and soul of the transformer model is a feed-forward neural network (Figure 1.6.1.1). These models typically take token-level embeddings as their input and try to predict the next token. For example, if the model is presented with the input *The boy went outside to fly his* _____, the large language model may assign high probabilities to the output tokens *kite* or *airplane*.[1]

In addition to a feed-forward neural network, the transformer architecture also includes a self-attention mechanism. Self-attention helps the model learn which words are related to each other, which has been a driving factor in the success of the transformer model over its predecessors. For

[1] Technically, the token-level of large language models is not analogous to words. For example, GPT-2's tokenizer tokenizes *kite* into two tokens: *k*, and *ite*. As such, GPT-2 assigns a high probability to *k* as the upcoming token in that context.

Figure 1.6.1.1: A visualization of a feed-forward neural network.



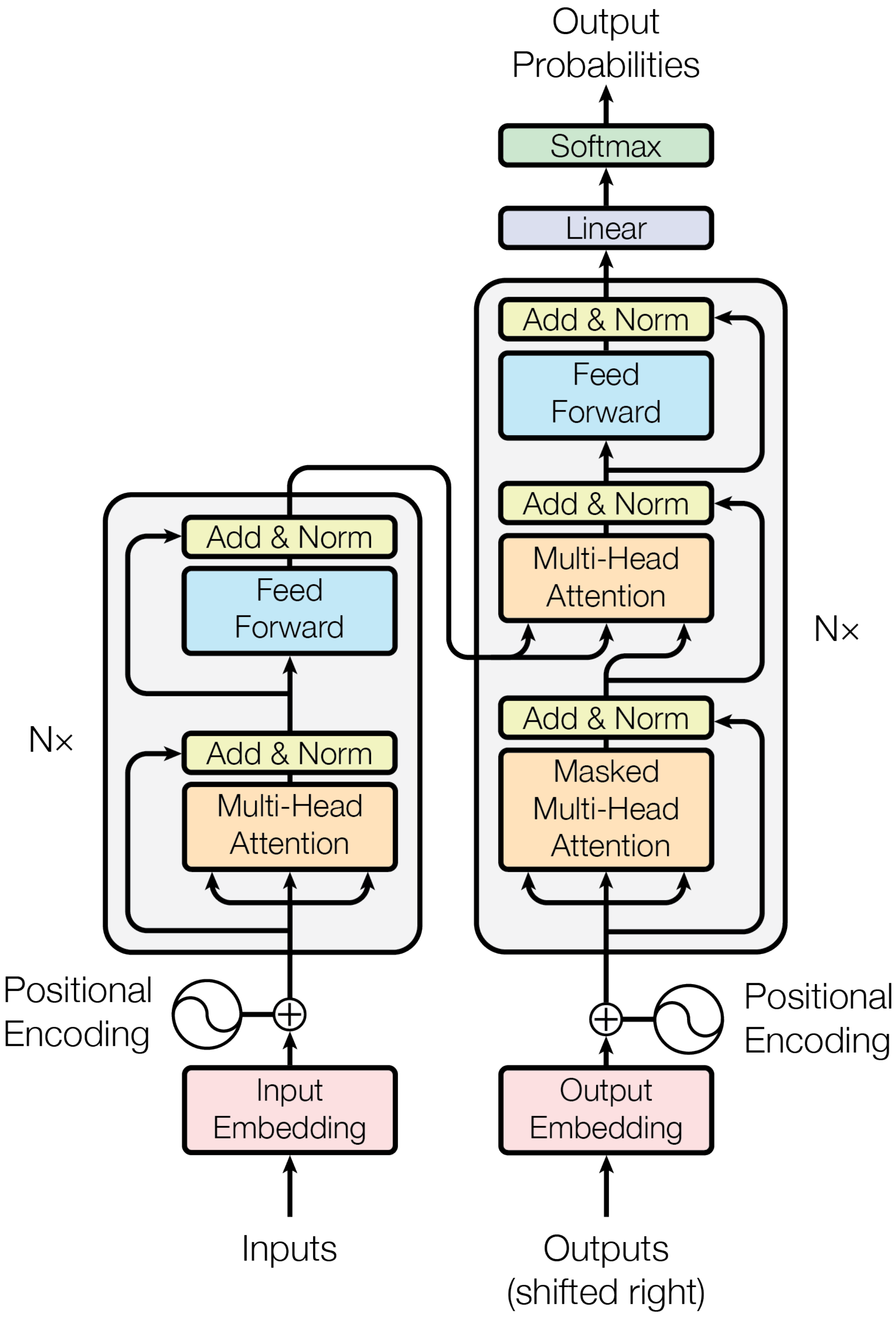
example, previous models such as Long-Short Term Memory (LSTM) models or Recurrent Neural Network (RNN) models struggled with long-term dependencies (Al-Selwi et al., 2023; Bengio et al., 1993). The self-attention mechanism was one of the solutions proposed to address this difficulty. The self-attention mechanism is a way to quantify which words are most related to each other. Specifically, the self-attention mechanism computes the strength of the relationship of each pair of words in the sentences (Vaswani et al., 2017). Thus in the sentence, *the students that go to school are bright*, the self-attention mechanism would assign a high value to the pair {*students, are*} because the word *students* is very relevant for predicting *are* (as opposed to *is*).

The full transformer model architecture is presented in Figure 1.6.1.2, reproduced from Vaswani et al. (2017).

### 1.6.2 Lessons from Transformer Models

Transformer models have achieved state-of-the-art performance on many benchmarks and are undoubtedly able to produce fluent, human-like text. However, it remains unclear to what extent

Figure 1.6.1.2: The transformer model architecture, reproduced from Vaswani et al. (2017).



they do this in a human-like manner. Specifically, how much are these models simply memorizing as opposed to learning something more abstract about the language?

On one hand, modern large language models are trained on trillions of tokens of data (Groeneveld et al., 2024). This is magnitudes larger than humans, who have heard an average of 350 million words by the time they enter college (Levy et al., 2012). Due to the size of the training data, there has been a lot of skepticism about how much they're actually learning and how much they're simply parroting from the training data (Bender et al., 2021). This skepticism is furthered by the fact that a lot of the training data for high-end large language models is either not open-access, or so huge that it is difficult to work with.

Further, there is evidence that large language models do copy a decent amount from their training. For example, Haley (2020) demonstrated that many of the BERT models are not able to reliably determine the correct plural form for novel words. Similarly, Li & Wisniewski (2021) demonstrated that BERT relies on memorization when producing the correct tense for a word as opposed to learning a more general linguistic pattern.

However, there is also a substantial amount of evidence that language models are learning more general patterns of the language (Lasri et al., 2022; Li et al., 2023; Li & Wisniewski, 2021; McCoy et al., 2023; Misra & Mahowald, 2024; Weissweiler et al., 2025; Yao et al., 2025). For example, Lasri et al. (2022) examined whether BERT is able to use the correct inflection of verbs in semantically incoherent contexts (e.g., *colorless green ideas sleep furiously*). They found that while BERT does do worse when the context is semantically incoherent, the decrease in performance is comparable to the decrease we see in humans.

Additionally, Li et al. (2023) examined BERT's performance on subject-verb and object-past agreements in French. They found that BERT uses abstract knowledge to predict subject-verb and object-past agreements.

There is also evidence that other transformer models can learn more abstract knowledge as well. For example, McCoy et al. (2023) examined the text that GPT-2 produces in relation to its training data. They found that while GPT-2 does copy a great deal, it also produces both novel words and syntactic structures.

More recently, there has been an interesting line of research examining language models trained on a more human amount of data. For example, Misra & Mahowald (2024) demonstrated that a language model trained on the BabyLM-strict corpus (Warstadt et al., 2023) (a corpus containing a comparable amount of data as humans receive) can learn article-adjective-numeral-noun constructions (AANNs). AANNs are constructions such as *a beautiful five days*. They occur relatively infrequently in English but humans still learn natural preferences for these. For example, while *a beautiful five days* is perfectly grammatical, *a blue five pencils* is not. They found that even after removing all AANN occurrences from the training data, the large language model is still able to learn these constructions. They further demonstrated that this is likely learned from similar constructions, such as *a few days*. The results of this show that even when trained on a comparable amount of data as humans, language models are still able to learn general patterns in the language.

Similarly, Yao et al. (2025) examined how language models learn the dative alternation. The dative alternation is a common construction in English where one can say either *give X to Y* or *give Y X*. Humans have preferences for these such as a length and an animacy preference. They trained a language model on a comparable amount of data to humans. Crucially they removed dative alternations that contained a length or animacy bias. They found that while the effect weakens, there is still an effect of length. They argued that this is evidence that language models are learning general patterns of the language. These results taken together with previous results demonstrate the ability of transformer models to learn general patterns in the language.

However, there is still a lot we don't know. What factors determine whether models learn general patterns of the language as opposed to relying on item-specific preferences? For example, humans seem to be sensitive to a combination of type and token frequency when they generalize beyond a specific word (Harmon & Kapatsinski, 2017). Are language models sensitive to similar factors? Further is this knowledge represented in a similar way as humans? That is, are the general preferences that large language models learn similar to those that humans learn? Understanding the answers to these questions is important for evaluating these models as theories of human language learning.

## 1.7 Outline of Dissertation

In the present dissertation, we provide an in depth examination of how humans trade off between storage and computation. In the next chapter, we examine whether predictability drives storage and how holistic representations are accessed. In Chapter 3, we examine how holistically stored items are represented. Chapter 4 examines how large language models trade off between storage and computation. Chapter 5 examines how stored items are represented in large language models. Chapter 6 examines whether storage accounts can explain frequency-dependent preference extremity. Finally, Chapter 7 examines a novel prediction found in Chapter 6.

# Chapter 2

# Does Predictability Drive the Holistic Storage of Compound Nouns?

## 2.1 Introduction

Learning a language is not a trivial task. In order to be successful, learners must accurately segment the continuous speech stream into smaller segments, including phrases, words, morphemes, and phonemes. One of the main questions that arises out of this task is what exactly is the size of the units that learners are storing? That is, are they storing individual words, entire sentences, phrases, or some combination of all of these? One possibility is that learners store very little outside of words and idioms. For example, traditional theories have argued that learners don't store any more than they need to: they store only what they can't form compositionally using a set of rules, and generate everything else (e.g., Chomsky, 1965). For example, inflected words, such as *walked* would be generated by accessing the stored root, *walk*, and then applying a past tense rule that generates *walked* from the root. Similarly, a phrase like *I don't know* would be generated by accessing each of the individually stored words *i*, *don't*, and *know*.

On the opposite side of this theoretical spectrum, another possibility is that learners store everything, including entire sentences. Ambridge (2020) argued for exactly this, specifically arguing that everything a learner hears "is stored with its meaning, as understood in that individual situation" and that unwitnessed novel-forms are produced using on-the-fly analogy across stored exemplars Ambridge (2020). For example, producing a novel plural form, like *wugs*, would consist of analogizing (on-the-fly) over multiple stored exemplars (e.g., *cats*, *chairs*, *dogs*, etc).

It is also possible that what gets stored is somewhere in between these two extremes. For example, usage-based construction grammar approaches have posited that a lot more than just words are stored – including high frequency phrases – but rather than storing everything, or storing only the most basic units, that storage is driven by usage (Arnon & Snider, 2010; R. H. Baayen et al., 2011; Bybee, 2003; Bybee & Hopper, 2001; Goldberg, 2003; Morgan & Levy, 2015b, 2016a, 2024; O'Donnell, 2016; Tomasello, 2005). That is to say, the size of the units stored is driven by the statistical distribution of the language that the learner is producing and perceiving. For example, Bybee (2003) drew an analogy to learning to play a piece on the piano:

> An important result of learning to play several pieces is that new pieces are then easier to master. Why is this? I hypothesize that the player can access bits of old stored pieces and incorporate them into new pieces. The part of a new piece that uses parts of a major scale is much easier to master if the player has practiced scales than is a part with a new melody that does not hearken back to common sequences. This means that snatches of motor sequences can be reused in new contexts. The more motor sequences stored, the greater ease with which the player can master a new piece.

In this same line of thinking, Bybee (2003) further argued against a strictly traditional view, stating that learning the English past tense *-ed* requires learning a series of words that contain that segment (e.g., *played*, *spilled*, *talked*) and that these are not necessarily flushed from memory after learning the English past tense marker.

There is no shortage of evidence for the holistic storage of multi-word phrases. For example, high-frequency phrases, such as *I don't know,* have been shown to undergo phonetic reduction that isn't seen in other low or mid-frequency phrases containing *don't* (Bybee & Scheibman, 1999) suggesting that the representation of *I don't know* is separate from the representation of each of the individual words. In other words, the susceptibility of high-frequency phrases to phonological change is strong evidence that they may come to have a mental representation for the whole expression (i.e., holistic storage). This example is not an outlier either, there are many examples of high-frequency phrases undergoing phonetic reduction: *going to*, want to*,* *have to*, etc (Bybee, 2003). In Korean, Yi (2002) demonstrated that in multi-word phrases containing the adnominal future marker (*-l*), the tensification of the consonant following the adnominal is predicted by the phrasal frequency. That is, in high-frequency phrases, consonants following the adnominal became tense at a higher rate than in low-frequency phrases.[1]

Evidence for holistic storage is not limited to phonological effects, either. In the Psycholinguistics literature, Siyanova-Chanturia et al. (2011) demonstrated that readers are sensitive to the ordering of binomials in English. In an eye-tracking experiment, participants read frequent binomial expressions in English in their preferred order (e.g., *bride and groom*) and their reversed order (*groom and bride*). They found that the preferred orderings were read faster. Further, Morgan & Levy (2016a) investigated whether the results from Siyanova-Chanturia et al. (2011) could be attributed to abstract knowledge of binomial orderings (e.g., a preference for male names before female names) or whether they were due to participants' direct experience with those items (e.g., hearing one ordering of a specific binomial more often than the complementary ordering). They developed a probabilistic model to approximate native English speakers' ordering preferences and combined that with a forced-choice and a self-paced reading task in order to investigate whether ordering preferences were driven by abstract knowledge or direct experience of the expression. They found that reading times for frequent binomials were influenced only by relative frequency (i.e., direct experience), not abstract knowledge. That is to say, ordering preferences of frequent binomials weren't explained by abstract ordering preferences, but rather by linguistic experience with the specific binomial, suggesting that high-frequency binomials are stored holistically.

Similarly, O'Donnell (2016) tested 4 probabilistic models on their ability to learn the English

[1] A similar effect has been demonstrated on the word-level as well in Korean, where epenthesis has been documented to occur more often in high-frequency words than in low-frequency words (Lee & Kapatsinski, 2015). This suggests that there may not be a clear division between the representation of high-frequency phrases and high-frequency words.

past tense and derivational morphology. Specifically, they tested a Full-parsing model, which stores minimal-sized units only, a Full-listing model, which stores the entirety of units only, an Exemplar-based model, which stores all units and all sub-units consistent with the data, and finally a Productivity as an Inference model, which, similar to the Exemplar-based Inference model, can store both smaller and larger structures, but probabilistically determines which items to store based on the data. They found that the Inference-based model performed the best overall for both past tenses and derivational morphology. In other words, storing units of varying sizes (as opposed to just minimal or maximal-sized units) seems to be the most conducive to learning the various morphological paradigms in English.

Despite the clear evidence for the holistic storage of some multi-word units, however, it is still largely unclear what determines whether a unit is stored holistically. For example, it is possible that storage is driven by either **phrasal frequency** (Bybee & Hopper, 2001) or by the mutual **predictability** of a phrase's component parts [i.e., how predictable the whole phrase is from part of the phrase; O'Donnell (2016)]. For example, as previously stated, there is an abundance of evidence that high-frequency phrases are more susceptible to phonetic reduction than low-frequency phrases (Bybee, 2003; Bybee & Scheibman, 1999). Additionally, high-frequency phrases have been shown to lose the recognizability of their component parts relative to low-frequency phrases (Kapatsinski & Radicke, 2009). For example, *up* is harder to recognize in *pick up* than in *run up*. On the other hand, in the learning literature, there is significant evidence that learning is driven by prediction error as opposed to raw co-occurrence statistics. For example, Ramscar et al. (2013) demonstrated that in word learning, children rely on more than simple co-occurrence statistics but also on how *informative* – that is, how *predictive* – a cue is of an outcome (relative to other cues). Specifically, they demonstrated that children rely on not only co-occurrence rate, but also background rate (how often a cue is present without an outcome). In other words, assuming doors have a higher co-occurrence rate and lower background rate than all the other competing cues (e.g., brown, house, room) for the word *door*, then children will learn that doors are the best predictor of the word *door* (Ramscar et al., 2013).

A similar debate persists in the speech perception literature, where Pierrehumbert (2001) argued that internal representations reflect the raw frequency distribution of the input. On the other hand, Olejarczuk et al. (2018) argued that the learning of phonemes is driven not by co-occurrence statistics (i.e., raw frequency), but rather by surprisal (i.e., prediction error). In other words, learners are actively predicting upcoming phonemes and update their beliefs in proportion to how surprising the upcoming phoneme is. Thus the debate between co-occurrence vs predictability in the role of learning is not unique to the word learning literature.

Additionally, if learners are storing more than just single-word units, what are the processing consequences of this? For example, as mentioned earlier, Kapatsinski & Radicke (2009) investigated the recognition of the particle *up* in phrases of varying frequencies and found that the recognition of the particle *up* is significantly more difficult in a high-frequency phrases than in low frequency phrases, suggesting that high frequency units 'fuse' together, losing some of the recognizability of their individual parts.

On the other hand, Staub et al. (2007) investigated the effects of plausibility on the reading times of familiar and novel compound nouns, which were compound nouns with high and low phrasal frequency respectively. Participants read sentences which contained a novel compound noun or a familiar compound noun (See the sentences below) in a plausible condition (a) or an implausible condition (b). Crucially, the second noun in the compound eliminated the local implausibility such that every sentence was plausible after reading the second noun. For example, in 1b *The zookeeper spread out the monkey...* is locally implausible, however upon reading the second noun in the compound, *medicine*, the local implausibility is eliminated.

1. **Novel Compound**

   **1a** The zookeeper picked up the monkey medicine that was in the enclosure.

   **1b** The zookeeper spread out the monkey medicine that was in the enclosure.

2. **Familiar Compound**

**2a** Jenny looked out on the huge mountain lion pacing in its cage.

**2b** Jenny heard the huge mountain lion pacing in its cage.

They found that the size of the plausibility effect was the same for both novel and familiar compound nouns. That is to say, while familiar items were read more quickly than novel items, and there was an increase in reading times in the implausible condition, the size of the plausibility effect was not different for familiar items (relative to novel items). However, if familiar items are stored holistically, one might expect that readers would predict the second noun upon reading the first, thus eliminating the local implausibility. Thus, if these items are stored holistically it begs the question of what the processing consequences of storage are. On the other hand, it may just be that these items are not stored. For example, it is possible that, as has been previewed throughout the introduction, phrasal frequency may not be the driving factor of storage and that it is actually predictability that might be driving storage. If this is the case, then it is possible that the reason for a lack of an interaction effect in Staub et al. (2007)'s results is due to their stimuli being low predictability compound nouns. For example, while *mountain lion* has a high phrasal frequency, *mountain* is not very predictable of *lion* (that is, the probability of *lion* following *mountain* is fairly low, despite the overall phrase having a relatively high frequency).

Thus there are two main problems that the present study aims to provide insight on: what exactly drives holistic storage, and what are the processing consequences of storage? In Experiment 1, we first replicate Staub et al. (2007)'s experiment using a maze task (Boyce et al., 2020). In Experiment 2, we use the same methodology, but instead of using high (phrasal) frequency compound nouns, we use high *predictability* compound nouns (e.g., *peanut butter*). By using the highest predictability compound nouns from the google *n*-grams corpus [Michel et al. (2011), we ask whether the difference in reaction times between the locally implausible and plausible contexts differs depending on whether the compound noun is highly predictable or not. For example, if highly predictable compound nouns are stored holistically, it is possible that when listeners hear, or read, the first noun in a highly predictable compound noun they may access the second noun as well and/or a holistic compound noun representation. If this is the case, then locally implausible contexts should not incur as much processing difficulty when the compound noun is highly predictable because the second noun eliminates the local implausibility. Lastly in Experiments 3 and 4 we replicate these with eye-tracking.

## 2.2 Experiment 1

### 2.2.1 Methods

#### Participants

Participants were presented with sentences online via ibex farm, a web-based experiment software platform that is freely-available (github.com/addrummond/ibex) and were recruited through the University of California Linguistics/Psychology Human Subjects Pool. To prevent selection bias, participants signed up for the experiment blindly, without knowledge of the content of the experiment. 146 participants were recruited, however 30 participants were excluded for having an overall accuracy below 70% (in this case, accuracy is operationalized as choosing the correct word; an inaccuracy would be choosing the ungrammatical distractor word), leaving a total of 116 participants. All participants self-reported being native English speakers.

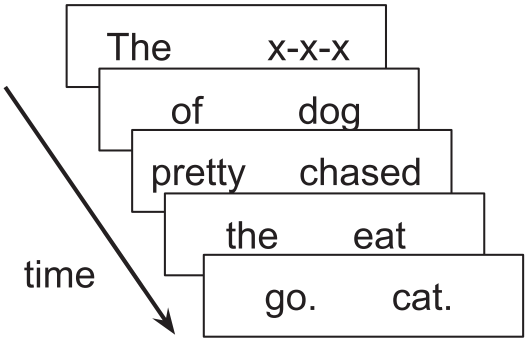
#### Stimuli

The experimental sentences were sentences containing compound nouns from (Staub et al., 2007) which varied upon two dimensions: local plausibility and familiarity. Locally plausible sentences were sentences in which the reading at the first noun was plausible and locally implausible sentences were sentences in which the reading at the first noun in the compound was implausible (see example sentences 1 and 2). Local plausibility was a within-item effect and familiarity (the frequency of the compound noun) was a between-item effect. Examples 1 and 2 above exemplify each condition: the first sentence in each is locally plausible while the second one is locally implausible. For example, in sentence 2a, it is semantically plausible that *Jenny looked out on the huge mountain...* but not semantically plausible that *Jenny heard the huge mountain* (2b). Altogether, our stimuli consisted of 24 novel items, 24 familiar items [taken from Staub et al. (2007)], and 188 filler sentences in order to avoid participants discerning the experimental design.

**Procedure**

Experiment 1 is a direct replication of Staub et al. (2007) using the A-Maze task (Boyce et al., 2020) instead of eye-tracking.[2] In the A-maze task, participants are presented with the first word in the sentence and then have to correctly choose between an ungrammatical distractor word and the next word in the sentence. When participants select the correct word, they continue to the next word in the sentence until the sentence is finished. The distractor words for the A-maze were generated automatically following Boyce et al. (2020) using the Gulordava model Gulordava et al. (n.d.). The locations of the distractor word and target word were counterbalanced so that they appeared an equal number of times on the left and right side of the screen. For each word, the reaction time was recorded along with whether the subject chose the correct item or not. See Figure 2.2.1.1 for a visualization of the maze task, reproduced from Boyce et al. (2020).

Figure 2.2.1.1: A visualization of the maze task, reproduced from Boyce et al. (2020).



Sentences were presented in a random order and each word was presented an equal number of times on the left and right side of the screen. Additionally, each item appeared an equal number of

[2] The maze task was used due to the limitations of the COVID-19 pandemic

times in the implausible and plausible context and no participant was presented with the same item in more than one condition. The complete dataset included 9994 response tokens.

### Analysis

The data was analyzed using Bayesian linear regression models, as implemented in the *brms* package (Bürkner, 2017) within the R programming environment R Core Team (2022). We subsetted the data into two sets based on the region: one set for the first noun in the compound noun and one set for the second noun in the compound. The primary dependent variable was log reaction time for both of these regions (following Boyce et al., 2020). The primary independent variables were plausibility and familiarity (following Staub et al., 2007). We modeled reaction time as a function of plausibility and familiarity, including their interaction, with maximal random effects (Barr et al., 2013). The formula used for the model is presented in equation Equation 2.1 below, with *Plaus* as plausibility and *Famil* as familiarity. All models were sum-coded

$$ReactionTime \sim Plaus * Famil + (Plaus * Famil|Subject) + (Plaus|Item) \tag{2.1}$$

### 2.2.2 Results

As mentioned in the methods section, for the purpose of the analysis, the data was divided into two regions: the N1 region and the N2 region, which were the first and second noun in the compound noun respectively. The results of the Bayesian regression model for the N1 region are presented in Table 2.1 and in Figure 2.2.2.1, and the results of the N2 region are presented in Table 2.2 and in Figure 2.2.2.2.

For the N1 region, there was an increase in reaction time for the implausible condition relative to the plausible condition. There was no such effect for familiarity. In other words, while participants took longer selecting the correct word in the implausible condition, their reaction times were not affected by the familiarity of the compound noun. This is expected given that the familiarity condition was not the frequency of the first noun, but rather the frequency of the compound noun as a whole. Additionally, there was no interaction effect between plausibility and familiarity.

At the N2 region, there was an increase in reaction time in the plausible condition and a decrease in reaction time in the familiar condition, but no interaction effect. In other words, participants were slower to choose the correct word in the plausible condition. They were also quicker to choose the correct word if the compound noun was familiar. However, plausibility did not mediate the effects of familiarity. That is to say, the size of the plausibility effect was not different for familiar versus novel compound nouns. Following Wagenmakers et al. (2010), a post-hoc Bayes factor analysis was conducted to compare the interaction effect to the null hypothesis (interaction effect = 0). We found a Bayes Factor value of 18.15 which constitutes strong support for the null hypothesis.

Table 2.1: Model results examining the effect of plausibility and frequency for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 6.822 | 0.023 | 6.777 | 6.866 | 100.00 |
| Plausibility | 0.060 | 0.010 | 0.040 | 0.080 | 83.86 |
| Familiarity | 0.015 | 0.015 | -0.014 | 0.044 | 100.00 |
| Plausibility:Familiarity | -0.003 | 0.010 | -0.023 | 0.017 | 38.76 |

Table 2.2: Model results examining the effect of plausibility and frequency for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 6.874 | 0.025 | 6.824 | 6.924 | 100.00 |
| Plausibility | -0.066 | 0.009 | -0.083 | -0.049 | 0.06 |
| Familiarity | -0.070 | 0.019 | -0.109 | -0.033 | 0.00 |
| Plausibility:Familiarity | 0.003 | 0.009 | -0.015 | 0.020 | 63.36 |

Figure 2.2.2.1: Plot of log reaction time at the N1 region as a function of plausibility and familiarity.

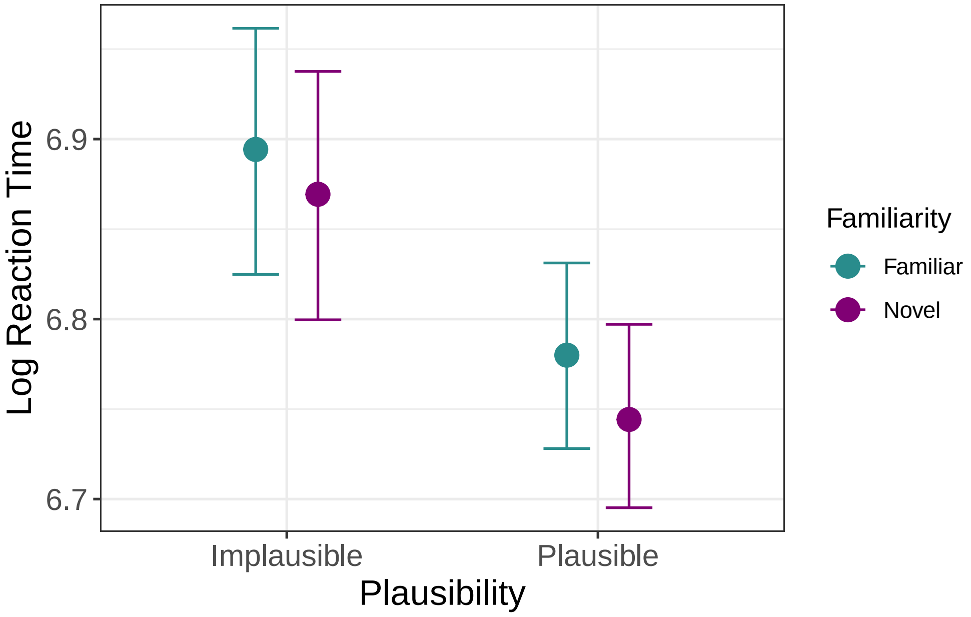
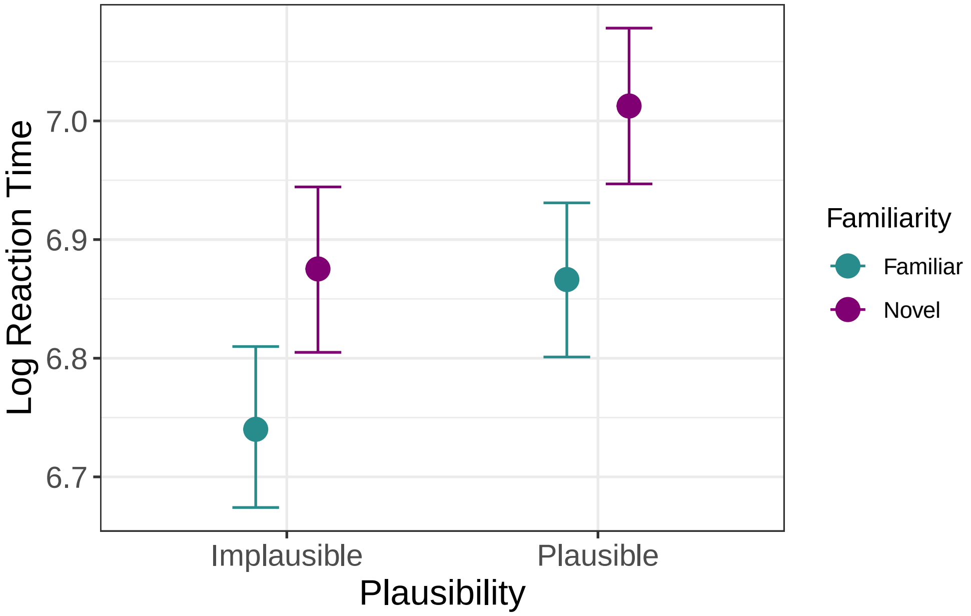


Figure 2.2.2.2: Plot of log reaction time at the N2 region as a function of plausibility and familiarity.



### 2.2.3 Discussion

Our results directly replicate Staub et al. (2007) using the Maze task, demonstrating the viability of this method for the tasks at hand. For the N1 region, while there was a clear increase in reaction time for items in the implausible condition, there was no interaction effect between plausibility and familiarity. In other words, the effect of plausibility was the same for both familiar and novel compound nouns. If familiar compound nouns are stored holistically, however, it is possible that we would see less of a (im)plausibility effect relative to novel items, because readers might be predicting the second noun in the compound upon reading the first noun. Recall Table 2, reproduced below for convenience:

1. **Novel Compound**

    **1a** The zookeeper picked up the monkey medicine that was in the enclosure.

    **1b** The zookeeper spread out the monkey medicine that was in the enclosure.

2. **Familiar Compound**

    **2a** Jenny looked out on the huge mountain lion pacing in its cage.

    **2b** Jenny heard the huge mountain lion pacing in its cage.

It is possible that if *mountain lion* was stored holistically, then upon reading *Jenny heard the huge mountain...*, the reader might have less difficulty with the local implausibility (relative to a low-frequency compound noun) because they would predict *lion*, which would eliminate the implausibility (*heard the mountain lion* is not implausible). However, we do not see this. Instead, the effect of plausibility is similar for both familiar and novel items. One possible explanation for these results is that the familiar phrases are not necessarily stored. Instead storage might be driven by predictability. If this is the case then it would explain why we do not see this effect in Staub et al. (2007) or in Experiment 1, especially since all of the items used in Staub et al. (2007) are low predictability compound nouns.

At the N2 region, the decrease in reaction time for familiarity is not surprising given that familiarity, as previously mentioned, was based on the frequency of the compound noun as a whole, however the increase in reaction time for the plausible condition is interesting, especially since the sentences were only locally implausible on the N1 region: the second noun in the compound always eliminated the local implausibility. It is possible this increase in reaction time is a garden path effect for committing to an interpretation of the sentence with the N1 and having to reanalyze the sentence. For example, when reading *Jenny looked upon the huge mountain...*, after reading *lion*, the reader may need to reanalyze the sentence, as the subject is not looking upon a mountain at all, but rather they are looking at a *mountain lion*. However, in the implausible condition participants may not fully commit to the interpretation since it is locally implausible, and thus may be waiting for a choice that eliminates the implausibility, thus explaining the absence of a similar slowdown in the implausible condition.

In Experiment 3, we examine whether readers can overcome this local implausibility for high-predictability items.

## 2.3 Experiment 2

### 2.3.1 Methods

#### Participants

Participant recruitment was identical to Experiment 1. 105 participants were recruited, and 19 participants were excluded for being below 70% accuracy, leaving a total of 86 participants. All participants self-reported being native English speakers.

#### Stimuli

We operationalized predictability through the odds ratio of the compound noun to the first word when that word is not followed by the second word in the compound noun which is exemplified in Equation 2.2.

$$\frac{\text{count}(\textit{peanut butter})}{\text{count}(\textit{peanut}) - \text{count}(\textit{peanut butter})} \tag{2.2}$$

In non-mathematical terms, Equation 2.2 quantifies how predictable the first noun is of the second noun (i.e., how likely the second noun is to follow after the first noun, relative to every other word that could follow). For example, the odds ratio of *peanut butter* would be the odds ratio of the compound noun – *peanut butter* – to the first noun – *peanut* – when *butter* does not follow it.

In order to collect the most predictable compound nouns, we searched the Google *n*-grams corpus (Michel et al., 2011) using the ZS Python package (Smith, 2014). We then collected the compound nouns with the highest predictability values, using the following exclusion criteria: excluding words with a match count below 90,000,[3] excluding nonsense words, proper nouns, technical words (e.g., *tenth circuit*), and words in which we could not create locally plausible and implausible sentences.[4] We gathered a total of 37 compound nouns for our high predictability condition. We subsequently normed the sentences we created using the high predictability compounds, as well as the sentences from Staub et al. (2007) which we confirmed were all low predictability compounds relative to our compound nouns.

We followed the same methodology as Staub et al. (2007) for our norming procedure: we provided participants with each item in four conditions (see below) and asked participants to rate each sentence on a 7-point Likert scale in terms of how well the last word fit in the sentence. No participant rated more than one version of each sentence.

### 3. Norming Conditions

[3] This was done in order to help filter out nonsense words (e.g., *teawhit head*) as well as eliminate words that had high predictability scores but were just a product of the corpus and unlikely to reflect the input of human learners (e.g., *broomwheat tea* which has a predictability score of 287 in the corpus).

[4] Given our methodology, we needed to be able to make sentences that were plausible and implausible using the same compound noun. This restriction meant we had to exclude words like *Parmesan cheese* where it would be impossible for the reading at the N1 region to be implausible without the reading of the compound noun also being implausible.

**3a** Jimmy picked up the peanut (plausible, through the first noun).

**3b** Jimmy picked up the peanut butter (plausible, through the second noun).

**3c** Jimmy spread out the peanut (implausible, through the first noun).

**3d** Jimmy spread out the peanut butter (implausible, through the second noun).

Finally, we excluded items in which the implausible sentence through the first noun was rated more or similarly well to the other conditions (i.e., the plausible sentence through the first noun, the plausible sentence through the second noun, and the implausible sentence through the second noun). It is important to note that due to our experimental design, the implausible sentence through the second noun is technically plausible at the second noun, because the second noun eliminates the local implausibility. Thus this condition should also receive a high rating, despite being the implausible condition. The mean values for each condition are as follows: plausible, through the first noun: 5.58 (sd = 0.78); plausible, through the second noun: 5.41 (sd = 0.71); implausible, through the first noun: 3.13 (0.63); implausible, through the second noun: 5.47 (sd = 0.82).

After norming, we selected sentences such that the difference in plausibility values between the plausible and implausible conditions were roughly the same for the high predictability and low predictability conditions. This was done to avoid conflating an interaction effect between predictability and plausibility with an item-specific effect. That is, if the plausibility effect was smaller for high predictability sentences relative to the low predictability sentences, then it would be impossible to tell if the interaction effect between predictability and plausibility is meaningful or just a product of our stimuli. The mean plausibility difference for the low predictability items was 2.47 and the mean plausibility difference for the high predictability items was 2.48. We confirmed that there was not a significant difference in plausibility values through a t-test (t = 0.0446, df = 39, p = 0.52). After accounting for this, we ended up with 21 high predictability and 21 low predictability items (which were taken from Staub et al., 2007), for a total of 42 items. Lastly, in order to avoid participants discerning the experimental design we also included 188 filler items.

### Procedure

Following Experiment 1, we used the A-maze task (Boyce et al., 2020) with automatically-generated distractor items (Gulordava et al., n.d.). Our dependent variable was reaction time and our independent variables were plausibility and predictability. We again used ibex farm to run our maze task. Sentences were presented in a random order and each word was presented an equal amount of times on the left and right side of the screen. Additionally, each item appeared an equal number of times in the implausible and plausible context and no participant was presented with the same item in more than one condition.

### Analysis

The data was analyzed using Bayesian linear regression models, as implemented in the *brms* package (Bürkner, 2017) within the R programming environment (R Core Team, 2022). We subsetted the data into two sets based on the region: one set for the first noun in the compound noun and one set for the second noun in the compound. The primary dependent variable was log reaction time for both of these regions (following Boyce et al., 2020). The independent variables were plausibility and predictability. Reaction time was modeled as a function of plausibility and predictability, along with their interaction, with maximal random effects (Barr et al., 2013). The formula used for the model is presented in Equation 2.3 below, with *Plaus* as plausibility and *Predic* as predictability.

$$ReactionTime \sim Plaus * Predict + (Plaus * Predict|Subject) + (Plaus|Item) \tag{2.3}$$

## 2.3.2 Results

As mentioned in the methods section, for the purpose of the analysis, the data was divided into two regions: the N1 region and the N2 region, which were the first and second noun in the compound noun respectively. The results of the Bayesian regression models for the N1 region are presented in Table 2.3 and Table 2.4, and visualized in Figure 2.3.2.1 and Figure 2.3.2.2. The results of the N2 region are presented in Table 2.5 and Table 2.6, and visualized in Figure 2.3.2.3 and Figure 2.3.2.4.

With regards to the N1 region, Table 2.3 presents the results of the analysis we ran with predictability as a binary predictor (high or low), while Table 2.4 presents the results of the analysis we ran with predictability as a continuous predictor (operationalized as the log odds ratio). Our results demonstrate that, similar to experiment 1, there was an increase in reaction time for the implausible condition, but no effect for predictability or the interaction between the two.

With regards to the N2 region, Table 2.5 presents the results of the analysis we ran with predictability as a binary predictor (high or low), while Table 2.6 presents the results of the analysis we ran with predictability as a continuous predictor (operationalized as the log odds ratio). Our results, as in Experiment 1, demonstrate an increase in reaction time in the plausible condition and and a decrease in reaction time in the high-predictability condition, but no interaction effect between plausibility and predictability. As in Experiment 1, we once again conducted a post-hoc Bayes factor analysis to compare the interaction effect to the null hypothesis (interaction effect = 0). We found a Bayes Factor value of 15.67 which constitutes strong support for the null hypothesis.

Figure 2.3.2.1 and Figure 2.3.2.3 provide visualizations of the analyses run with predictability as a binary variable while Figure 2.3.2.2 and Figure 2.3.2.4 present analyses with predictability as a continuous variable.

Table 2.3: Regression analysis results for the N1 region with predictability as a binary predictor (high or low).

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 6.879 | 0.029 | 6.823 | 6.936 | 100.00 |
| Plausibility | 0.068 | 0.014 | 0.041 | 0.095 | 100.00 |
| Predictability | 0.034 | 0.024 | -0.012 | 0.081 | 92.30 |

| | | | | | |
|---|---|---|---|---|---|
| Plausibility:Predictability | -0.001 | 0.013 | -0.027 | 0.025 | 46.72 |

Table 2.4: Regression analysis results for the N1 region with predictability as a continuous predictor (log odds ratio).

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 6.876 | 0.030 | 6.817 | 6.935 | 100.00 |
| Plausibility | 0.069 | 0.013 | 0.042 | 0.095 | 100.00 |
| LogOdds | 0.005 | 0.007 | -0.009 | 0.020 | 76.50 |
| Plausibility:LogOdds | 0.001 | 0.004 | -0.007 | 0.009 | 61.64 |

Table 2.5: Regression analysis results for the N2 region with predictability as a binary predictor (high or low).

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 6.793 | 0.029 | 6.738 | 6.852 | 100.00 |
| Plausibility | -0.072 | 0.012 | -0.097 | -0.049 | 0.00 |
| Predictability | -0.080 | 0.026 | -0.132 | -0.030 | 0.12 |
| Plausibility:Predictability | 0.005 | 0.012 | -0.017 | 0.028 | 67.46 |

Table 2.6: Regression analysis results for the N2 region with predictability as a continuous predictor (log odds ratio).

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 6.802 | 0.026 | 6.752 | 6.853 | 100.00 |
| Plausibility | -0.073 | 0.012 | -0.097 | -0.048 | 0.00 |
| LogOdds | -0.033 | 0.007 | -0.046 | -0.019 | 0.00 |
| Plausibility:LogOdds | 0.001 | 0.004 | -0.006 | 0.008 | 60.44 |

Figure 2.3.2.1: Plot of the N1 region with predictability as a binary variable (high or low).

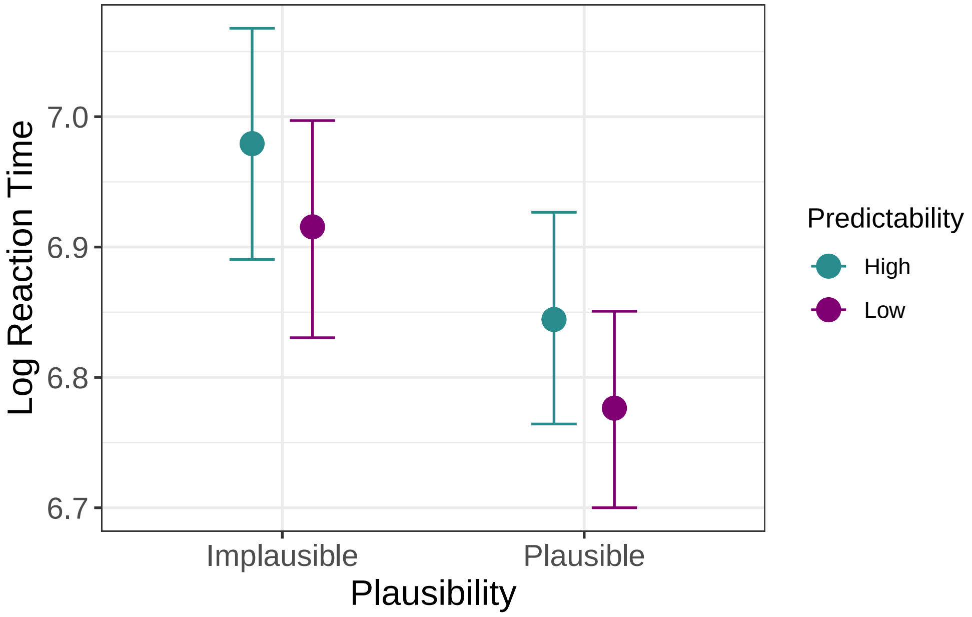


Figure 2.3.2.2: Plot of the N1 region with predictability as a continuous variable.

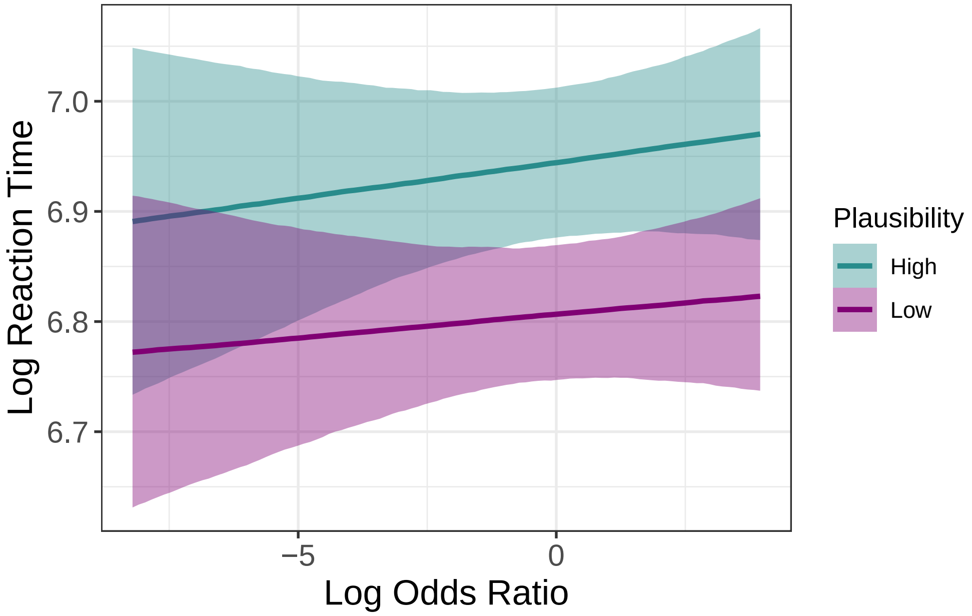


Figure 2.3.2.3: Plot of the N2 region with predictability as a binary variable (high or low).

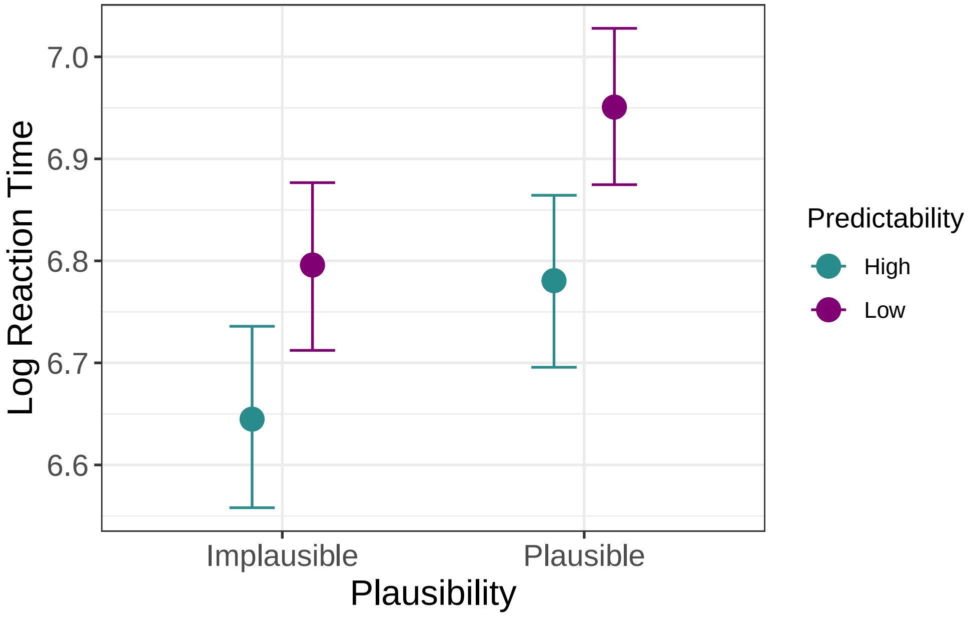
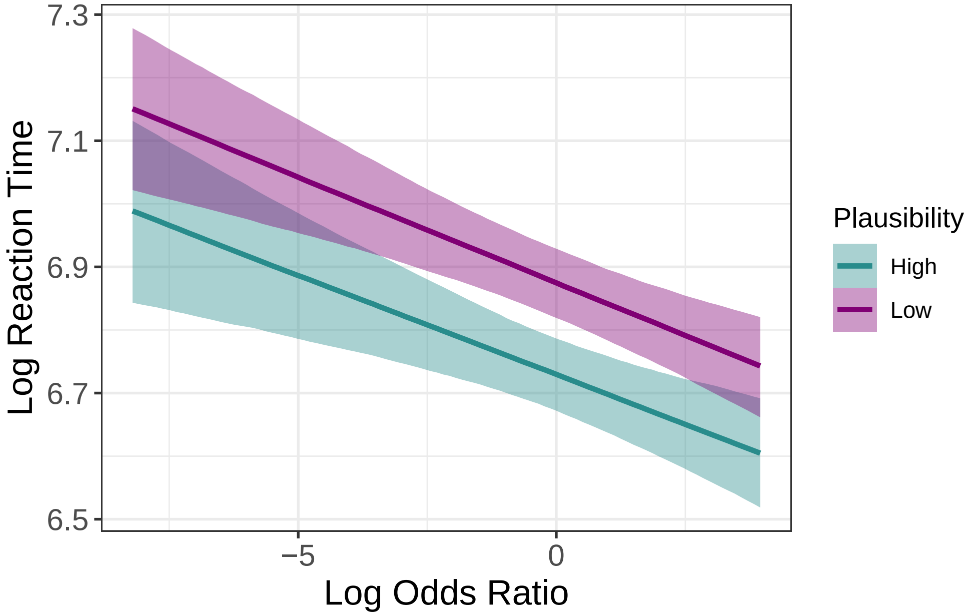


Figure 2.3.2.4: Plot of the N2 region with predictability as a continuous variable.



### 2.3.3 Discussion

Experiment 2 replicates and extends Experiment 1 using predictability instead of familiarity (i.e., phrasal frequency). Interestingly, the results of Experiment 2 were extremely similar to the results of Experiment 1: There was no interaction effect between predictability and plausibility on the RTs for the N1 condition. Additionally, while we see an effect of implausibility on the N1 region, we don't see an effect of predictability. This is expected since predictability is defined as the odds that the N2 appears given the N1, so we should see this effect on the N2 region, not the N1 region.

The results of the N2 region also bear remarkable similarities to our results in Experiment 1: There was a plausibility and predictability effect, but no interaction between the two. Specifically, there was an *increase* in reaction time for items in the plausible condition relative to the implausible condition. It is possible that, as mentioned in the discussion section of Experiment 1, this increase in reaction time is a garden path effect for committing to an interpretation of the sentence with the N1 and having to reanalyze the sentence.

## 2.4 Experiment 3

In Experiment 3, we directly replicate Experiment 1 using eye-tracking.

### 2.4.1 Methods

#### Participants

46 native English speakers were recruited from the University of California, Davis subjects pool. They were given course credit in exchange for their participation. All participants had normal or corrected vision.

### Materials

The materials were identical to Experiment 1.

We recorded participants' eye movements using the Eyelink 1000 Plus. We recorded pupil movements from the right eye. Participants were seated 850mm away from the screen. Our screen resolution was 1920x1080, 531.3mm in width, and 298.8mm in height.

Comprehension was checked for non-experimental trials and participants below 80% accuracy were excluded. Out of our 56 participants, 0 were excluded for falling below the accuracy threshold.

### Analyses

Prior to our analyses, sentences with blinks were excluded and fixations less than 80ms in duration and within one character of the nearest fixation were merged into that fixation (following Staub et al., 2007). For our regions of interest (the first noun and the second noun in the compound noun), we computed first fixation duration, first pass time, go-past time, and first-pass regression.

For each analysis, our independent variables were plausibility (high or low) and familiarity (high or low) and their interaction. We also included random slopes for condition and predictability by subject and plausibility by compound noun as well as intercepts for subject and compound noun. For each of our models, we used sum-coding, where the intercept represents the grand mean and the fixed-effect coefficient estimates represent the distance from the grand mean.

## 2.4.2 Results

### First Fixation Times

### N1

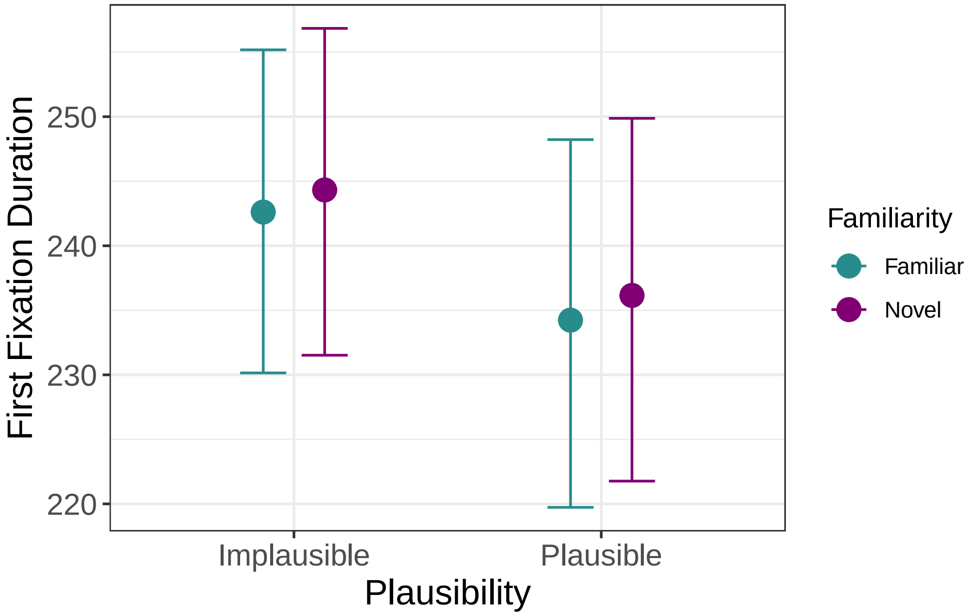
Our results for the effects of plausibility and familiarity on first fixation times are described in Table 2.7 and visualized in Figure 2.4.2.1.

We find an effect of plausibility, however we find no effect for familiarity and no interaction effect. The results suggest that regardless of familiarity, participants' first fixations are longer in the implausible context.

Table 2.7: Model results examining the effect of plausibility and frequency on first fixation times for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 239.234 | 5.536 | 228.463 | 250.049 | 100.000 |
| Plausibility | 4.188 | 2.156 | -0.022 | 8.378 | 97.400 |
| Familiarity | -0.866 | 2.399 | -5.589 | 3.784 | 35.525 |
| Plausibility:Familiarity | 0.027 | 2.398 | -4.843 | 4.782 | 49.675 |

Figure 2.4.2.1: Visualization of the effects of plausibility and familiarity on first fixation times for the N1 region.



### N2

Our results for the effects of plausibility and familiarity on first fixation times on the N2 region are described in Table 2.8 and visualized in Figure 2.4.2.2.

We find an effect of plausibility, however we find no effect for familiarity and no interaction effect. The results suggest that regardless of familiarity, participants' first fixations are longer in the implausible context.

Our results suggest an effect of familiarity such that familiar items had shorter first fixation times. We find no effect of plausibility which is expected because the N2 has no difference in terms of plausibility. We also find no interaction effect. These results suggest that readers' first fixation times on the N2 region are shorter for the familiar compounds.

Table 2.8: Model results examining the effect of plausibility and familiarity on first fixation times for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 249.615 | 6.468 | 236.382 | 262.346 | 100.000 |
| Plausibility | -2.094 | 2.652 | -7.400 | 3.120 | 21.575 |
| Familiarity | -6.973 | 4.270 | -15.430 | 1.443 | 5.425 |
| Plausibility:Familiarity | -1.055 | 2.365 | -5.669 | 3.620 | 32.475 |

### Gaze/First-Pass Time

### N1

Our results for the effects of plausibility and familiarity on gaze/first-pass times on the N1 region are described in Table 2.9 and visualized in Figure 2.4.2.3.

We find no effect of plausibility or familiarity on reading times for the N1 region. These results suggest that readers' gaze/first-pass times were not sensitive to the plausibility manipulation.

Figure 2.4.2.2: Visualization of the effects of plausibility and familiarity on first fixation times for the N2 region.

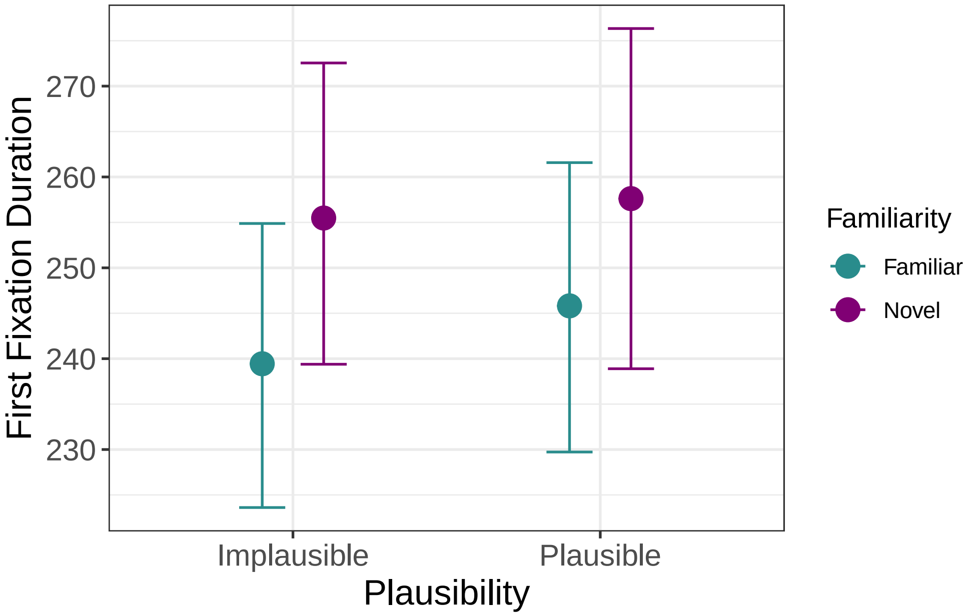
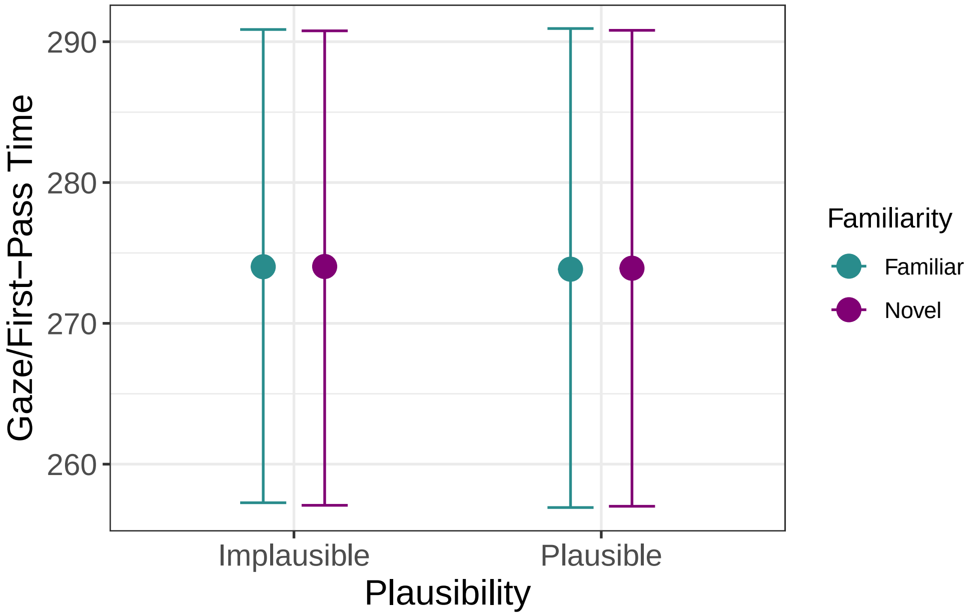


Table 2.9: Model results examining the effect of plausibility and familiarity on gaze/first-pass times for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 273.963 | 8.562 | 257.093 | 290.927 | 100.000 |
| Plausibility | 0.037 | 0.195 | -0.334 | 0.422 | 57.825 |
| Familiarity | -0.003 | 0.199 | -0.399 | 0.381 | 49.725 |
| Plausibility:Familiarity | 0.009 | 0.199 | -0.376 | 0.404 | 51.925 |

Figure 2.4.2.3: Visualization of the effects of plausibility and familiarity on Gaze/first-pass times for the N1 region.



N2

Our results for the effects of plausibility and familiarity on gaze/first-pass times on the N2 region are described in Table 2.10 and visualized in Figure 2.4.2.4.

We find no effect of plausibility, but we do find an effect of familiarity on reading times for the N2 region. These results suggest that readers' gaze/first-pass times on the N2 region were shorter for familiar compounds.

Table 2.10: Model results examining the effect of plausibility and familiarity on gaze/first-pass times for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 275.485 | 9.039 | 257.696 | 293.098 | 100.0000 |
| Plausibility | 0.048 | 3.755 | -7.501 | 7.445 | 50.5500 |
| Familiarity | -12.956 | 6.366 | -24.981 | -0.202 | 2.3250 |

Table 2.10: Model results examining the effect of plausibility and familiarity on gaze/first-pass times for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Plausibility:Familiarity | -3.879 | 3.775 | -11.309 | 3.425 | 15.2375 |

### Go-Past Time

### N1

Our results for the effects of plausibility and familiarity on go-past times in the N1 region are described in Table 2.11 and visualized in Figure 2.4.2.5.

Our results suggest a main-effect of plausibility as well as an interaction effect between plausibility and familiarity such that the increase in go-past times for familiar items is much greater than the increase is for novel items.

Table 2.11: Model results examining the effect of plausibility and familiarity on go-past times for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 363.001 | 18.576 | 325.948 | 399.737 | 100.00000 |
| Plausibility | 16.903 | 6.907 | 3.608 | 30.536 | 99.36667 |
| Familiarity | -3.748 | 7.888 | -19.444 | 11.847 | 31.34000 |
| Plausibility:Familiarity | 14.839 | 7.419 | 0.450 | 29.409 | 97.76667 |

### N2

Our results for the effects of plausibility and familiarity on go-past times in the N2 region are described in Table 2.12 and visualized in Figure 2.4.2.6.

Figure 2.4.2.4: Visualization of the effects of plausibility and familiarity on gaze/first-pass times for the N2 region.

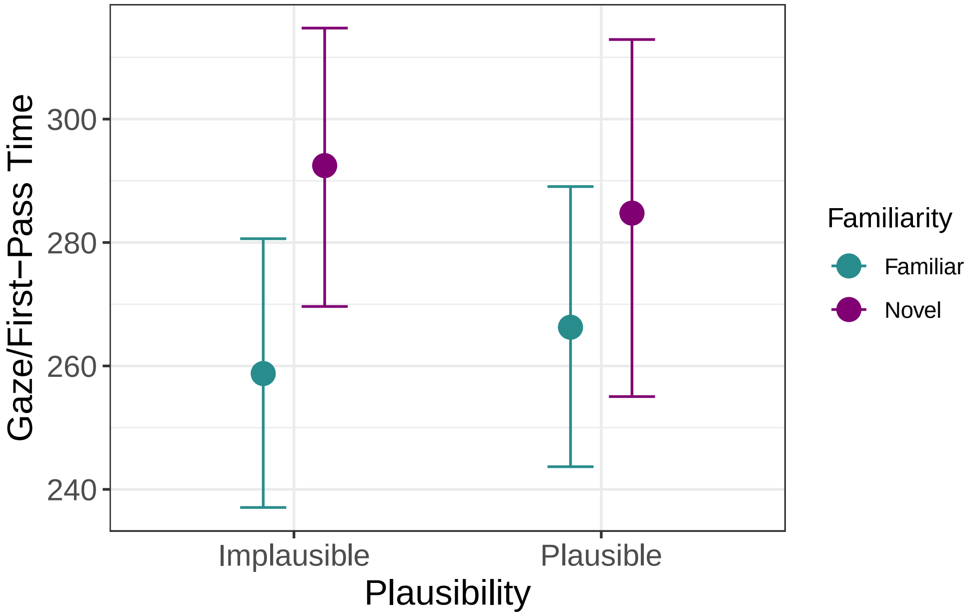
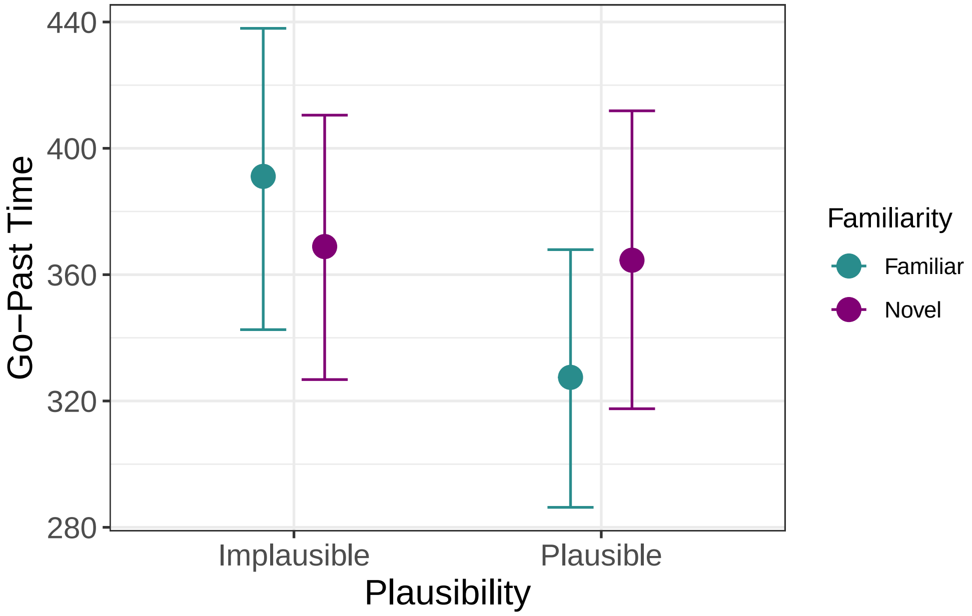


Figure 2.4.2.5: Visualization of the effect of plausibility and Familiarity on go-past times for the N1 region.

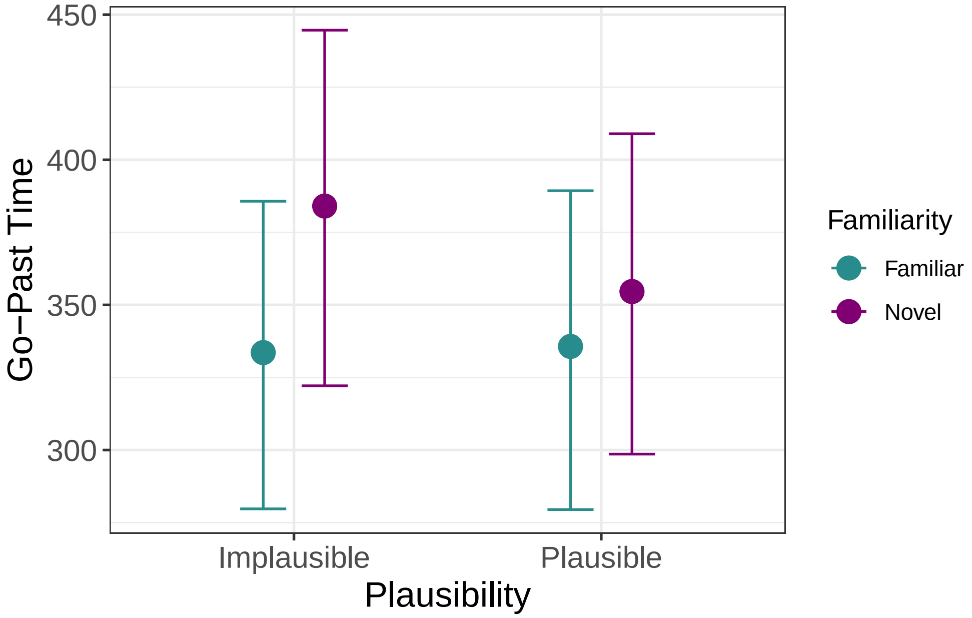


Our results suggest an effect of familiarity such that familiar items had shorter go-past times than novel items.

Table 2.12: Model results examining the effect of plausibility and familiarity on go-past times for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 351.825 | 21.876 | 308.493 | 394.204 | 100.000000 |
| Plausibility | 6.952 | 8.405 | -9.270 | 23.774 | 79.873333 |
| Familiarity | -17.433 | 13.362 | -44.885 | 7.492 | 8.966667 |
| Plausibility:Familiarity | -7.957 | 9.345 | -26.618 | 10.247 | 19.426667 |

Figure 2.4.2.6: Visualization of the effect of plausibility and familiarity on go-past times for the N2 region.



**First-Pass Regression**

**N1**

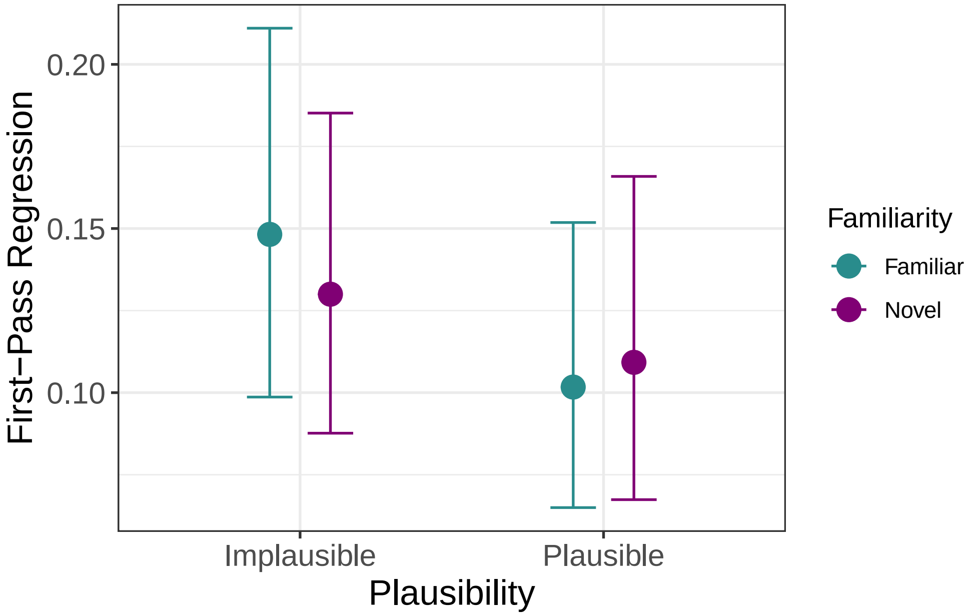
Our results for the effects of plausibility and familiarity on first-pass regressions in the N1 region are described in Table 2.21 and visualized in Figure 2.5.2.7.

Our results suggest an effect of plausibility on first pass regressions for the N1 region such that implausible items had more first-pass regressions than plausible items.

Table 2.13: Model results examining the effect of plausibility and familiarity on first-pass regression for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | -1.986 | 0.180 | -2.347 | -1.642 | 0.000 |
| Plausibility | 0.157 | 0.089 | -0.014 | 0.332 | 96.375 |
| Familiarity | 0.017 | 0.088 | -0.152 | 0.188 | 57.800 |
| Plausibility:Familiarity | 0.056 | 0.089 | -0.121 | 0.232 | 74.500 |

Figure 2.4.2.7: Visualization of the effect of plausibility and familiarity on first-pass regression for the N1 region.



**N2**

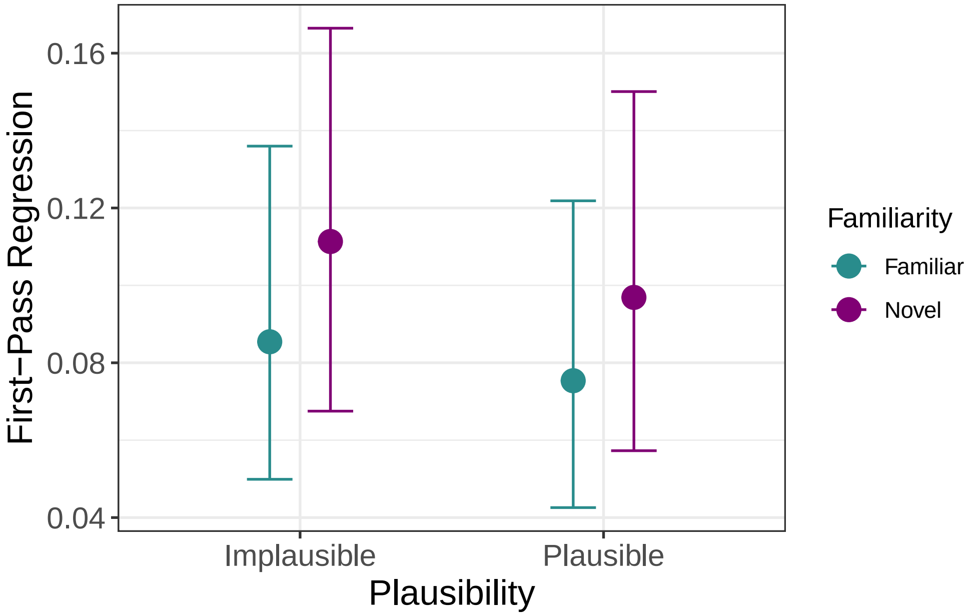
Our results for the effects of plausibility and familiarity on first-pass regressions in the N2 region are described in Table 2.22 and visualized in Figure 2.5.2.8.

Our results suggest an effect of familiarity on first pass-regressions for the N2 region suggesting that familiar items had fewer first-pass regressions than novel items.

Table 2.14: Model results examining the effect of plausibility and familiarity on first-pass regression for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | -2.307 | 0.212 | -2.748 | -1.910 | 0.000 |
| Plausibility | 0.074 | 0.092 | -0.103 | 0.256 | 78.450 |
| Familiarity | -0.140 | 0.118 | -0.371 | 0.096 | 11.575 |
| Plausibility:Familiarity | -0.004 | 0.088 | -0.179 | 0.168 | 48.575 |

Figure 2.4.2.8: Visualization of the effect of plausibility and familiarity on first-pass regression for the N2 region.



### 2.4.3 Discussion

Experiment 3 demonstrates that readers have longer first-fixation times, longer go-past times, and more first-pass regressions in implausible contexts than in plausible contexts. Further, we find an interaction effect in the opposite direction from predicted for go-past times: the effect of plausibility is greater for familiar items than in novel items. We will expand on this result in the Conclusion section.

## 2.5 Experiment 4

In Experiment 4, we replicate the results we found in Experiment 2 using eye-tracking.

### 2.5.1 Methods

#### Participants

56 native English speakers were recruited from the University of California, Davis subjects pool. They were given course credit in exchange for their participation. All participants had normal or corrected vision.

#### Materials

The materials here were identical to those in Experiment 2.

#### Procedure

We recorded participants' eye movements using the Eyelink 1000 Plus. We recorded pupil movements from the right eye. Participants were seated 850mm away from the screen. Our screen resolution was 1920x1080, 531.3mm in width, and 298.8mm in height.

Comprehension was checked for non-experimental trials and participants below 80% accuracy were excluded. Out of our 56 participants, 0 were excluded for falling below the accuracy threshold.

### Analyses

Prior to our analyses, sentences with blinks were excluded and fixations less than 80ms in duration and within one character of the nearest fixation were merged into that fixation (following Staub et al., 2007). For our regions of interest (the first noun and the second noun in the compound noun), we computed first fixation duration, first pass time, go-past time, and first-pass regression.

For each analysis, our independent variables were plausibility (high or low) and (log) predictability (high or low) and their interaction. We also included random slopes for condition and predictability by subject and plausibility by compound noun as well as intercepts for subject and compound noun. For each of our models, we used sum-coding, where the intercept represents the grand mean and the fixed-effect coefficient estimates represent the distance from the grand mean.

## 2.5.2 Results

### First Fixation Times

### N1

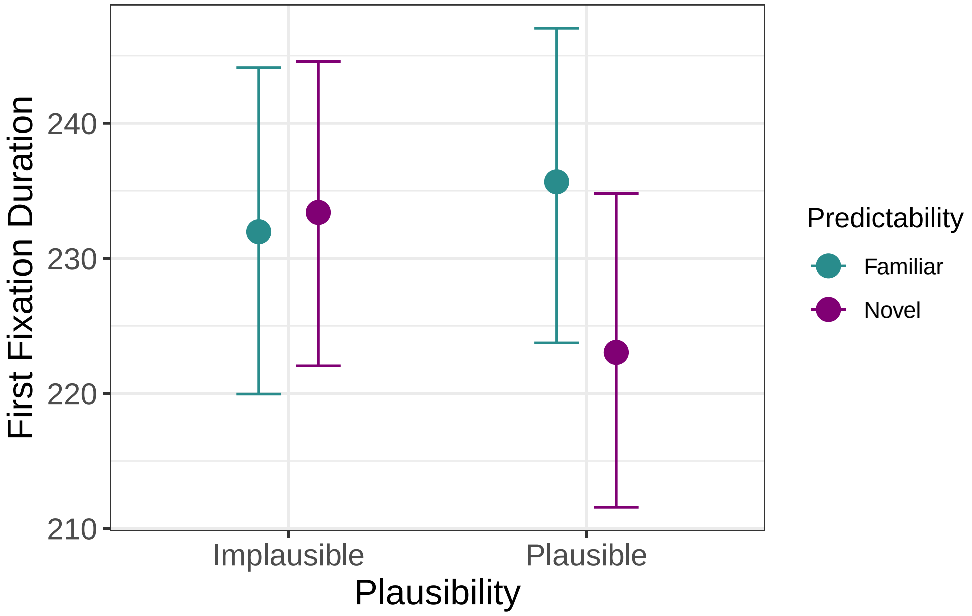
First, we examined the effects of plausibility and predictability on first fixation times for the first noun. Note that in our case, since plausibility was coded as -1 for plausible and 1 for implausible, a positive coefficient estimate of plausibility corresponds to longer fixation times for implausible items. Additionally, following Houghton et al. (2024) we also report the percentage of posterior samples greater than zero. Our model results can be found below in Table 2.15 and a visualization can be found in Figure 2.5.2.1.

Our results for first-fixation times, while non-significant, do show an interesting trend, with the effect of plausibility in the expected direction (although with only ~78% samples greater than zero). Further, we found a small but meaningful interaction effect such that the readers' first-fixation times of low-predictability items was affected more by the local implausibility than the high-predictability items were.

Table 2.15: Model results examining the effect of plausibility and predictability on first fixation times for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 231.016 | 4.433 | 222.356 | 239.502 | 100.00 |
| Plausibility | 1.661 | 2.020 | -2.289 | 5.695 | 78.95 |
| Predictability | 2.782 | 2.870 | -2.804 | 8.532 | 83.75 |
| Plausibility:Predictability | -3.473 | 2.012 | -7.418 | 0.522 | 4.15 |

Figure 2.5.2.1: Visualization of the effects of plausibility and predictability on first fixation times for the N1 region.



N2

Our results for the effects of plausibility and predictability on first fixation times in the N2 region are described in Table 2.16 and visualized in Figure 2.5.2.2

Our results demonstrate an effect of predictability on the first fixation times of the second noun in the compound noun. This is rather unsurprising because in the present study predictability is a measure of how expected the second noun is given the first. There also is not much of a plausibility effect on the N2, which is also not particularly surprising because the second noun eliminates the local implausibility making both sentences plausible at the N2 region.

Table 2.16: Model results examining the effect of plausibility and predictability on first fixation times for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 234.448 | 4.864 | 224.479 | 243.767 | 100.000 |
| Plausibility | -2.151 | 2.973 | -8.132 | 3.685 | 23.225 |
| Predictability | -4.136 | 2.988 | -9.887 | 1.887 | 8.575 |
| Plausibility:Predictability | -2.928 | 3.012 | -8.920 | 2.967 | 16.075 |

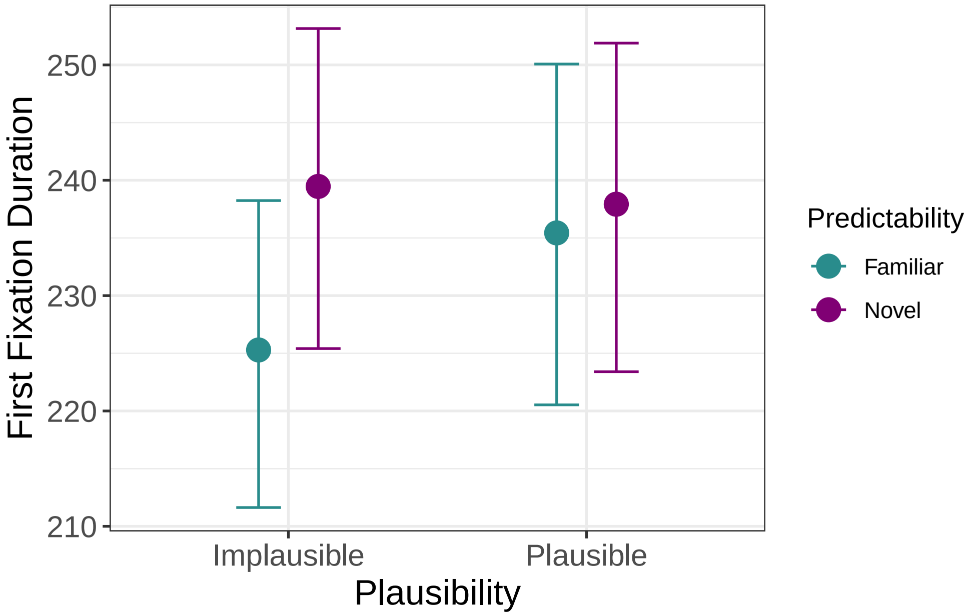
### Gaze/First-Pass Time

### N1

Our results for the effects of plausibility and predictability on gaze/first-pass times for the N1 region are presented in Table 2.17 and visualized in Figure 2.5.2.3.

Our results demonstrate no effect of either plausibility or predictability on the gaze times at the N1 region. We further examined our filler items to rule out an error with the eye-tracker. Our filler items contain a frequency manipulation and an analysis of the filler items demonstrated that the frequency manipulation did effect gaze times in our filler items, suggesting that the results here are

Figure 2.5.2.2: Visualization of the effects of plausibility and predictability on first fixation times for the N2 region.

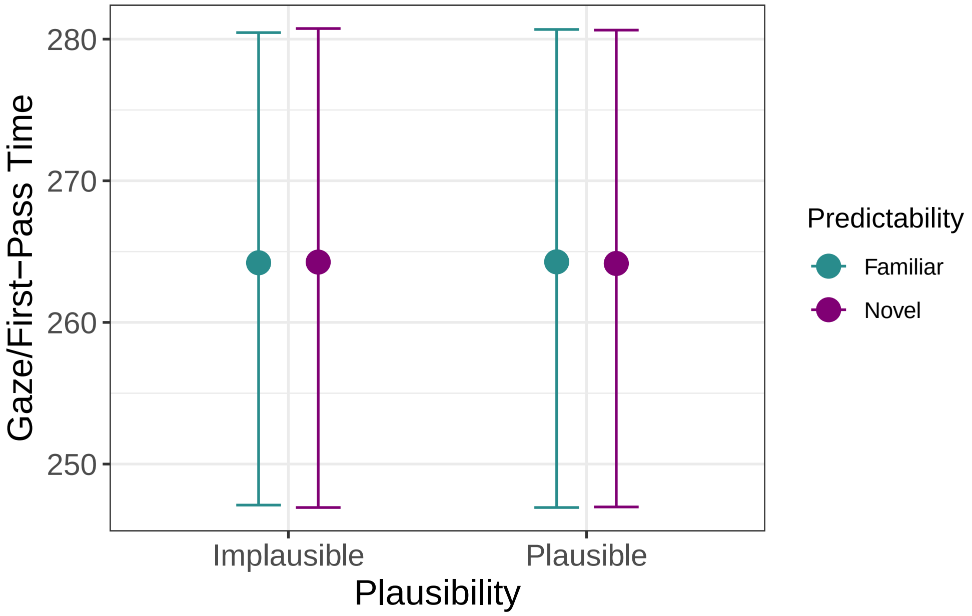


not due to any malfunction of the eye-tracker or cleaning procedure. We report the effects found in the filler items at the end of this section.

Table 2.17: Model results examining the effect of plausibility and predictability on Gaze/first-pass times for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 264.141 | 8.422 | 246.898 | 280.601 | 100.000 |
| Plausibility | -0.001 | 0.203 | -0.385 | 0.413 | 49.075 |
| Predictability | 0.005 | 0.198 | -0.394 | 0.387 | 51.475 |
| Plausibility:Predictability | -0.010 | 0.202 | -0.411 | 0.382 | 48.100 |

Figure 2.5.2.3: Visualization of the effects of plausibility and predictability on Gaze/first-pass times for the N1 region.



**N2**

Our results for the effects of plausibility and predictability on gaze/first-pass times for the N2 region are presented in Table 2.17 and visualized in Figure 2.5.2.3.

Our results demonstrate no effect of either plausibility or predictability on the gaze times at the N2 region. This is surprising because the effect of predictability should facilitate reading at the N2.

Table 2.18: Model results examining the effect of plausibility and predictability on Gaze/first-pass times for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 253.757 | 6.071 | 241.726 | 265.767 | 100.000 |
| Plausibility | -0.003 | 0.100 | -0.196 | 0.192 | 48.825 |
| Predictability | -0.003 | 0.101 | -0.207 | 0.190 | 48.625 |

Table 2.18: Model results examining the effect of plausibility and predictability on Gaze/first-pass times for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Plausibility:Predictability | 0.003 | 0.100 | -0.199 | 0.202 | 51.625 |

**Go-Past Time**

**N1**

Our results for the effects of plausibility and predictability on go-past times for the N1 region are presented in Table 2.19 and visualized in Figure 2.5.2.5.

The results for go-past times show no effect of predictability and plausibility on the N1 region.

Table 2.19: Model results examining the effect of plausibility and predictability on go-past times for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 357.207 | 16.452 | 325.205 | 388.957 | 100.00000 |
| Plausibility | 0.018 | 0.197 | -0.367 | 0.400 | 54.05000 |
| Predictability | 0.005 | 0.200 | -0.390 | 0.394 | 50.10000 |
| Plausibility:Predictability | -0.000 | 0.200 | -0.392 | 0.393 | 49.76667 |

**N2**

Our results for the effects of plausibility and predictability on go-past times for the N2 region are presented in Table 2.20 and visualized in Figure 2.5.2.6.

Figure 2.5.2.4: Visualization of the effects of plausibility and predictability on Gaze/first-pass times for the N2 region.

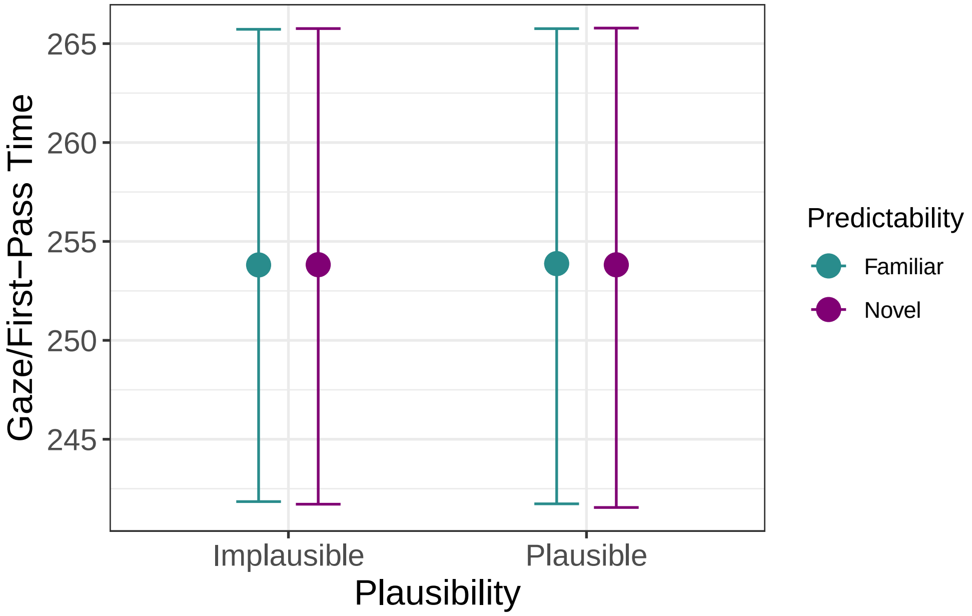
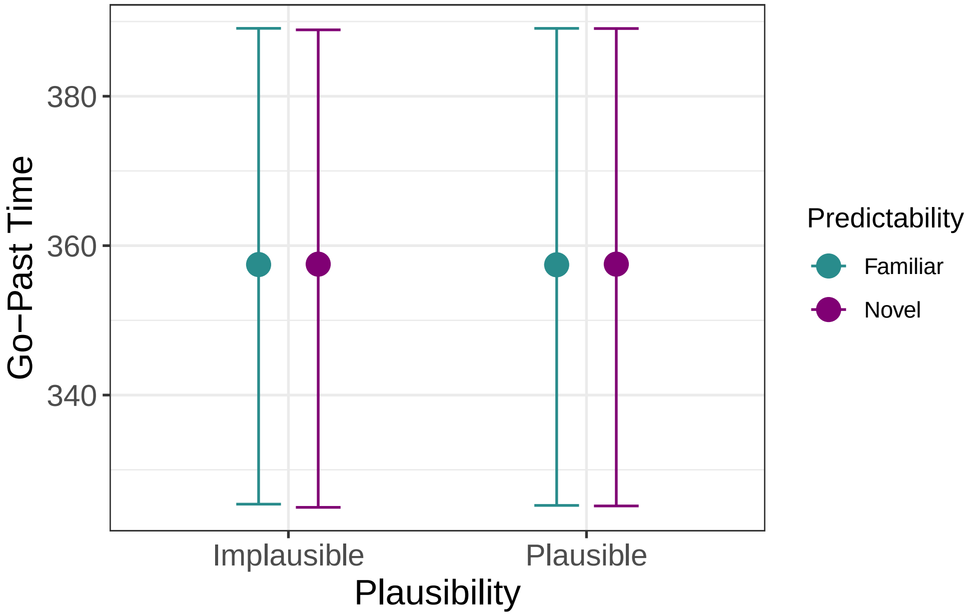


Figure 2.5.2.5: Visualization of the effect of plausibility and predictability on go-past times for the N1 region.

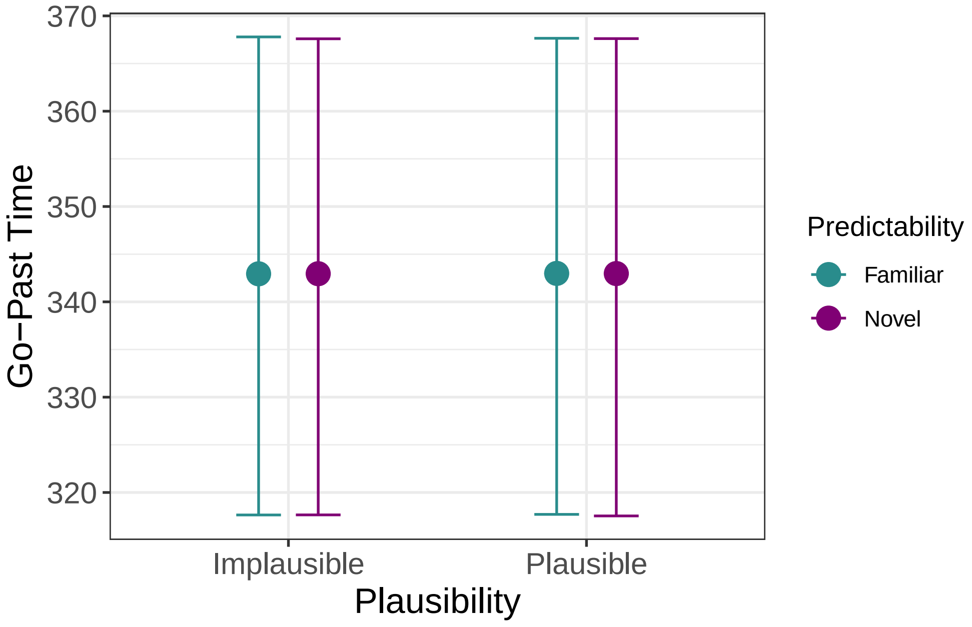


Our results for the N2 region similarly show no effect of predictability and plausibility on go-past times.

Table 2.20: Model results examining the effect of plausibility and predictability on go-past times for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 342.737 | 12.866 | 317.685 | 367.649 | 100.000 |
| Plausibility | -0.005 | 0.099 | -0.202 | 0.195 | 47.600 |
| Predictability | -0.001 | 0.100 | -0.198 | 0.195 | 50.000 |
| Plausibility:Predictability | -0.002 | 0.100 | -0.192 | 0.198 | 49.025 |

Figure 2.5.2.6: Visualization of the effect of plausibility and predictability on go-past times for the N2 region.



**First-Pass Regression**

**N1**

Our results for the effects of predictability and plausibility on the first-pass regressions on the N1 region are presented in Table 2.21 and visualized in Figure 2.5.2.7.

Our results suggest that readers are more likely to regress after the first fixation in the implausible condition compared to the plausible condition. Further, this plausibility effect is larger for high-predictability items than low predictability items. This is surprising because predictability is a measure of the N2, not the N1 and if readers are anticipating the N2 then it should alleviate the local implausibility at the N1 (which would result in a negative interaction effect, i.e. the opposite trend from what we see here).

Table 2.21: Model results examining the effect of plausibility and predictability on first-pass regression for the N1 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | -1.645 | 0.151 | -1.948 | -1.358 | 0.000 |
| Plausibility | 0.199 | 0.080 | 0.041 | 0.357 | 99.375 |
| Predictability | -0.049 | 0.086 | -0.214 | 0.123 | 27.750 |
| Plausibility:Predictability | 0.128 | 0.075 | -0.019 | 0.275 | 96.025 |

**N2**

Our results for the effects of predictability and plausibility on first-pass regressions in the N1 region are presented in Table 2.22 and visualized in Figure 2.5.2.8.

Our results no effect of predictability or plausibility. In other words, unexpectedly high-predictability items had similar first-pass regressions as low-predictability items in the N2 region.

Figure 2.5.2.7: Visualization of the effect of plausibility and predictability on first-pass regression for the N1 region.

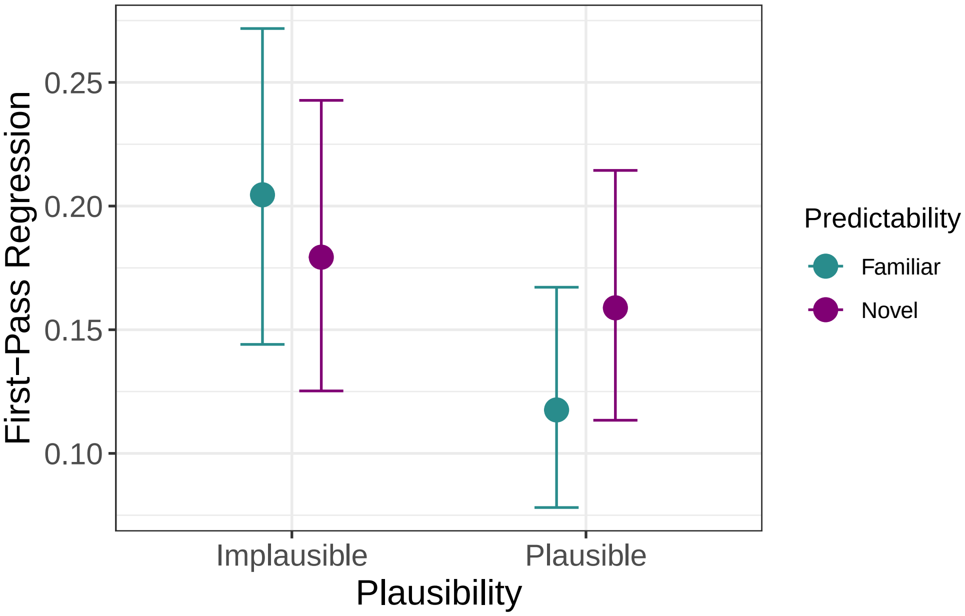
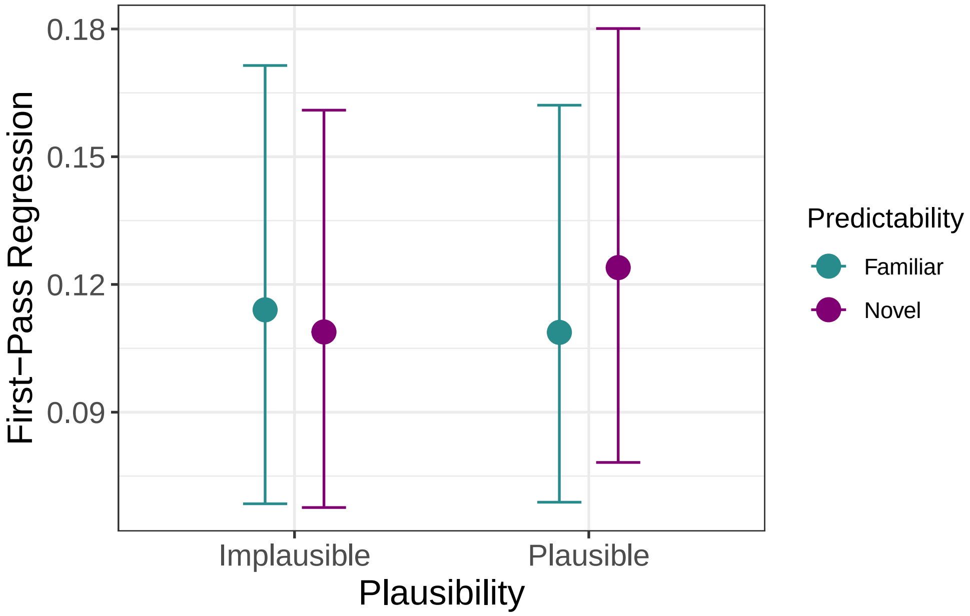


Table 2.22: Model results examining the effect of plausibility and predictability on first-pass regression for the N2 region.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | -2.061 | 0.176 | -2.426 | -1.728 | 0.000 |
| Plausibility | -0.026 | 0.106 | -0.241 | 0.184 | 40.625 |
| Predictability | -0.022 | 0.102 | -0.223 | 0.177 | 41.100 |
| Plausibility:Predictability | 0.049 | 0.097 | -0.136 | 0.239 | 69.975 |

Figure 2.5.2.8: Visualization of the effect of plausibility and predictability on first-pass regression for the N2 region.



### Filler Items

In order to confirm that our results weren't due to measurement error, we used examined our filler items to sanity check our findings. Our filler items contained a frequency manipulation where some items were low-frequency and some were high-frequency. The Psycholinguistics literature has demonstrated the robustness of the effects of frequency on reading measures, thus we examined our filler items to confirm that that the results for Gaze/First-Pass times and Go-Past time were not due to measurement error.

### First Fixation Times

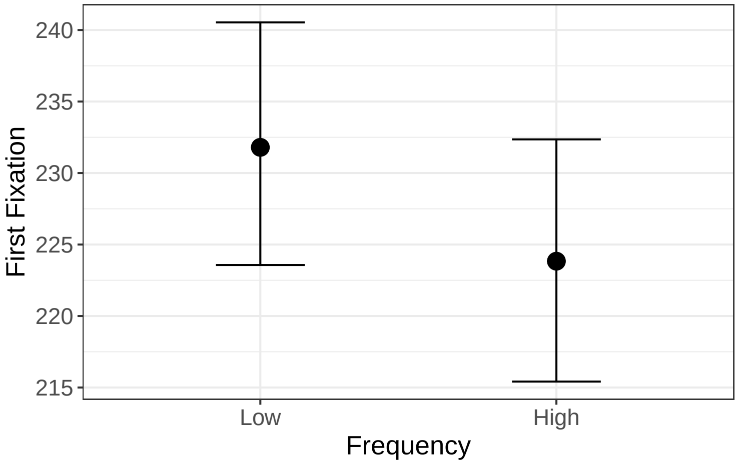
Our results for the effects of frequency on first fixation times for our filler items is presented in Table 2.23 and visualized in Figure 2.5.2.9.

Our results, unsurprisingly, demonstrate longer first fixation times for low frequency items compared to high frequency items.

Table 2.23: Model results examining the effect of frequency on first fixation times in our filler materials.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 227.843 | 3.942 | 220.186 | 235.404 | 100.0 |
| Frequency | 4.011 | 1.848 | 0.414 | 7.642 | 98.5 |

Figure 2.5.2.9: Visualization of the effect of frequency on first fixation times in our filler materials.



### Gaze/First-Pass Time

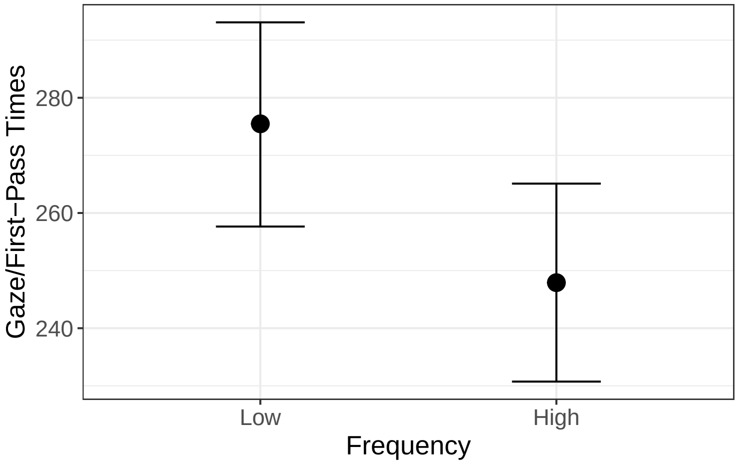
Our results for the effects of frequency on gaze/first-pass times for our filler items is presented in Table 2.24 and visualized in Figure 2.5.2.10.

Similarly, our results demonstrate longer gaze times for low-frequency items compared to high-frequency ones. This suggests that the results we found in Experiment 4 are not due to measurement error.

Table 2.24: Model results examining the effect of frequency on gaze/first-pass times in our filler materials.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 261.710 | 7.805 | 246.338 | 276.992 | 100.00000 |
| Frequency | 13.825 | 4.371 | 5.179 | 22.322 | 99.88333 |

Figure 2.5.2.10: Visualization of the effect of frequency on gaze/first-pass times in our filler materials.



### Go-Past Time

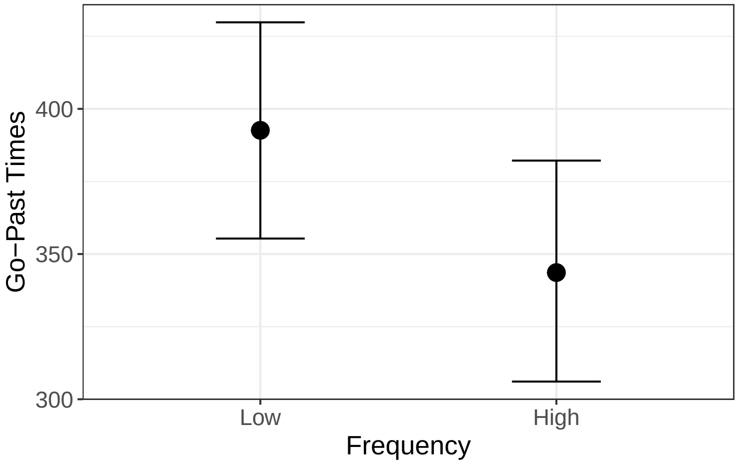
Our results for the effects of frequency on go-past times for our filler items is presented in Table 2.25 and visualized in Figure 2.5.2.11.

For go-past times as well, we do find an effect such that low-frequency items had longer go-past times. These results further confirm that the results found in Experiment 4 were not due to measurement error.

Table 2.25: Model results examining the effect of frequency on go-past times in our filler materials.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | 368.199 | 17.340 | 334.329 | 402.836 | 100.0 |
| Frequency | 24.391 | 7.863 | 9.001 | 39.509 | 99.8 |

Figure 2.5.2.11: Visualization of the effect of frequency on go-past times in our filler materials.



### First-Pass Regression

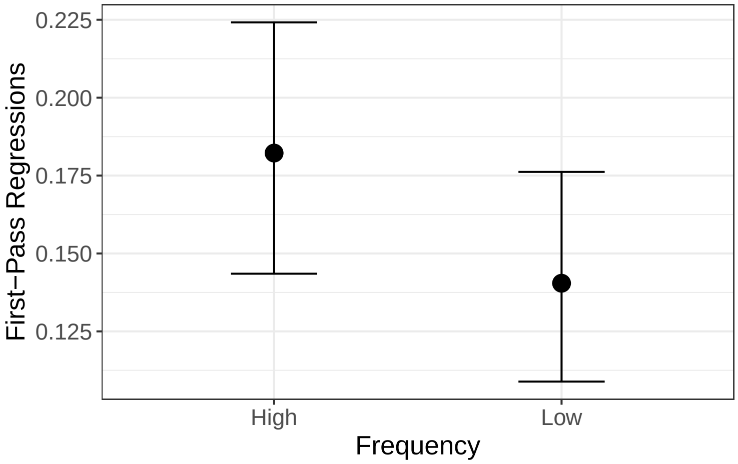
Finally, our results for the first-pass regression rates of our filler items is presented in Table 2.26 and visualized in Figure 2.5.2.12.

The results of first-pass regressions demonstrate a greater rate of first-pass regressions for low frequency items relative to high frequency items. These results taken together suggest that the effects of plausibility and predictability on each of the eye-tracking measures was not a consequence of measurement error.

Table 2.26: Model results examining the effect of frequency on first-pass regressions in our filler materials.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | -1.659 | 0.129 | -1.921 | -1.417 | 0.0 |
| Frequency | 0.156 | 0.058 | 0.044 | 0.271 | 99.6 |

Figure 2.5.2.12: Visualization of the effect of frequency on first-pass regressions in our filler materials.



### 2.5.3 Discussion

The results in Experiment 4 confirmed an effect of plausibility in first-fixation times for low-predictability items, but not for high-predictability items. This is difficult to reconcile with the first-pass regressions where the opposite pattern is observed: there is an effect of plausibility on first-pass regressions for high-predictability items but not low-predictability items.

We also find no effect of plausibility for gaze duration or go-past times, which was unexpected. Further, upon inspecting our filler items we were able to confirm that this was not a consequence of measurement error.

## 2.6 Conclusion

The present studies examined the processing of compound nouns in locally implausible and locally plausible contexts, specifically with respect to their phrasal frequency and predictability. In Experiment 1 we replicated Staub et al. (2007) using the A-maze task (Boyce et al., 2020) and found an increase in reaction time for the implausible condition at N1 region, but no interaction effect between plausibility and familiarity. Additionally at the N2 region, we found an increase in reaction time for the plausible condition relative to the implausible condition and a decrease in reaction time for high predictability items relative to low predictability items.

In Experiment 2 we extended Experiment 1 using predictability as the key measure instead of phrasal frequency. Similar to Experiment 1, we found an increase in reaction time for the implausible condition at the N1 region, but again found no interaction effect between plausibility and predictability. Also similar to Experiment 1, we found an increase in reaction time at the N2 region for the plausible condition and a decrease in reaction time for the high predictability items.

In Experiments 3 and 4 we replicated the two experiments with eye-tracking. In Experiment 3, we find an effect of plausibility in first fixation times, go-past times, and first-pass regressions. We also find an interaction effect in go-past times such that high-frequency items had higher go-past times in the implausible condition, but low-frequency items did not. In Experiment 4, we find an effect of plausibility in first-fixation times for low-predictability items (but not high-predictability items) and an an effect of plausibility in first-pass regressions for high-predictability items but not low-predictability items.

Overall the results of experiments 1 and 2 suggest that the predictability of the second noun in the compound (given the first noun) has very little facilitatory effect on the processing of the first noun. Importantly, the increase in reaction time in the implausible condition for the N1 region was not mediated by the predictability of the compound noun. If participants were predicting the second noun upon reading the first noun, then we might expect to have seen a decrease in reaction time for the high-predictable items in the implausible condition relative to the low-predictability items.

Experiments 3 and 4 provide mixed-evidence. On one hand, there seems to be a general effect of implausibility, however in some reading measures such as first-fixation times, predictability seemed to alleviate this effect. On the other hand, for other measures, implausible contexts actually affected familiar/high-predictability items more than the novel/low-predictability items.

There are a few possible explanations for the results we found. One possibility is simply that our high-predictability compound nouns aren't stored holistically. It is important to note that our compound nouns were the most predictable compound nouns in the entire Google *n*-grams corpus, thus it seems unlikely that they weren't predictable enough to be stored, though it may be that English compound nouns have relatively low predictability relative to other multi-word phrases. Instead it may be possible that predictability isn't the driving force of storage.

Another possibility is that the high-predictability compound nouns are stored holistically, but the processing consequences of them being stored holistically are such that there is no facilitatory effect in the processing of the first noun in the compound noun. This would certainly beg the question, however, of what exactly the processing consequences of holistic storage are. Perhaps the primary advantages to holistic storage are in the domain of production, rather than processing.

Finally, it is possible that there may be competing forces. On the one hand, high-predictability items may help overcome the local implausibility (as evidenced by the alleviation in first-fixation times found in Experiment 4). On the other hand, there's also evidence from the literature that people are generally more acceptable of low-frequency words in novel contexts than high-frequency words (Harmon & Kapatsinski, 2017). It is possible that these competing forces explain the mixed-results in the eye-tracking experiments.

Finally, with respect to the increase in reaction time at the N2 region in the plausible condition that we found in Experiments 1 and 2, we do not see this effect in Experiments 3 and 4, suggesting that this effect may be a task-specific effect. This result suggests that in the maze task, participants have a bias to analyze the first noun as the head noun and then have to reinterpret the sentence once it is clear that the noun is not the head noun. Further, since we don't see the same effect in the implausible condition, participants may not fully commit to an interpretation that is implausible.

In summary, the present study contributes to the current theories of sentence processing by demonstrating that predictability may not be the driving factor behind holistic storage, however given the lack of research demonstrating the specific processing consequences of holistic storage, it is possible that rather than predictability not driving holistic storage, either our task doesn't elicit a measurable effect of holistic storage, or holistic storage of a compound noun just doesn't facilitate the processing of the first noun in compound nouns.

# Chapter 3

# The effects of frequency and predictability on the recognition of *up* in English verb+up collocations.

## 3.1 Introduction

When a listener hears the phrase *trick or treat*, do they process it compositionally, processing each word individually before combining them into a single parse? Or do they access a single holistically stored representation of the phrase from memory? This question of to what extent larger-than-word constructions can be stored and accessed holistically is one that psycholinguists have been interested in for quite some time (Bybee, 2003; e.g., Bybee & Hopper, 2001; Goldberg, 2003; Nooteboom et al., 2002; Stemberger & MacWhinney, 1986, 2004).

Throughout the years different theories have argued for different degrees of holistic storage, with two theories in particular dominating the field. On one hand, Chomskyan theories (e.g., Chomsky, 1965; Pinker & Ullman, 2002) have proposed that only necessary items (e.g., items that can't be formed compositionally) are stored.[1] On the other hand, usage-based theories (e.g., Bybee, 2003) have proposed that many items that could in principle be formed compositionally can be stored under certain usage-based conditions, such as frequency of use.

[1] Although some theories (e.g., Pinker & Ullman, 2002) have accepted that some very high-frequency items may be stored due to human memory, but these theories are much more conservative about what is stored compared to usage-based theories.

Traditional Chomskyan theories (e.g., Chomsky, 1965; Pinker & Ullman, 2002) have argued that processing multi-word phrases is completely compositional: each piece is accessed individually and then combined to form the larger meaning. Some exceptions are reserved for idioms and other outliers, which can't be formed compositionally. More specifically, Chomskyan views of storage argue that whether an item is stored is determined purely by the degree of compositionality. According to these theories, if a multi-word expression can be composed from its parts then there is no need to holistically store the expression, and thus it is not stored holistically. For example since *I don't know* can be processed compositionally, it would be processed by composing a representation from each of the individual words, *I, don't,* and *know*. On the other hand, *kicked the bucket* would be stored holistically because there's very little relationship between the meaning of the individual words and the meaning of the expression (i.e., it's non-compositional).

Chomskyan theories of storage gained popularity partly because storage was thought to be a valuable resource that was taken up only by units that necessitated storage. This was perhaps influenced by the limited storage space of sophisticated computers at the time. In recent times, however, we've learned that the brain may have dramatically more space for storage than we had previously realized, with an upper bound of $10^{8432}$ bits (Wang et al., 2003). This is magnitudes larger than any current estimate of how much storage language requires.[2] Considering this, it might not come as a surprise that there has been a rise in support for usage-based theories of holistic storage over the past few decades (Ambridge, 2020; H. Baayen et al., 2002; Bybee, 2003; Bybee & Hopper, 2001; Bybee & Scheibman, 1999; Kapatsinski, 2018; Kapatsinski & Radicke, 2009; Morgan & Levy, 2016a; Stemberger & MacWhinney, 1986, 2004; Zang et al., 2024).

Usage-based theories posit that more than just non-compositional items (e.g., multi-word expressions) may be stored holistically in the lexicon, arguing that storage is driven by usage-based factors. For example, factors like frequency or predictability of the phrase may influence whether the phrase is stored holistically or not. According to these theories, in addition to idioms and non-

[2] Indeed, Mollica & Piantadosi (2019) estimated that, in terms of linguistic information, humans store only somewhere between one million and ten million bits of information, meaning that even their upper estimate is well within the capacity of the brain.

compositional items, multi-word phrases such as *I don't know* may also be stored holistically if they are used frequently enough (e.g., Ambridge, 2020; Arnon & Snider, 2010; Kapatsinski, 2018; Kapatsinski & Radicke, 2009; Lee & Kapatsinski, 2015; Morgan & Levy, 2016a; Stemberger & MacWhinney, 1986, 2004; Tomasello, 2005).

While it has become a dominant view in the field that at least some multi-word items are stored, it remains unclear what exactly the size of the units being stored is and, more so, what the factors driving storage are. Further, if multi-word representations are stored holistically, what are the consequences of this in terms of language processing?

### 3.1.1 Evidence of Holistic Storage

There is no shortage of evidence for holistic multi-word storage (e.g., Bybee & Scheibman, 1999; Christiansen & Arnon, 2017; Stemberger & MacWhinney, 1986, 2004; Zwitserlood, 2018), especially in the phonology literature. For example, Bybee & Scheibman (1999) demonstrated that the word *don't* is reduced to a larger extent in the phrase *I don't know* than in other phrases containing *don't*. In other words, the phrase *I don't know* seems to have its own mental representation. If it was the case that the representation of *don't* in *I don't know* was the same as the representation of *don't* in other contexts, then one would expect *don't* to be equally reduced in both cases (which is contrary to the finding in Bybee & Scheibman, 1999). Similarly, in Korean, certain consonants undergo tensification when they occur after the future marker *-l*. The rate of this tensification is higher in high-frequency phrases than low-frequency phrases, further suggesting that high-frequency phrases may be stored holistically (Yi, 2002).

In addition to the phonology literature, the Psycholinguistics literature has also provided an abundance of evidence for multi-word storage. For example, Siyanova-Chanturia et al. (2011) demonstrated that binomial phrases (e.g., *cat and dog*) are read faster in their more frequent ordering than in their less frequent ordering. Further, in a follow-up study, Morgan & Levy (2016a) demonstrated that these ordering preferences for frequent binomials are not due to abstract ordering preferences (e.g., a preference for short words before long words), but are rather driven by experience with the specific binomial (i.e., how frequent each binomial ordering is), providing additional evidence that frequent phrases are stored holistically.

Similarly, Arnon & Snider (2010) demonstrated that frequent multi-word phrases are read faster than lower frequency multi-word phrases, even after accounting for the frequency of the individual words. This suggests that humans are sensitive to the frequencies of multi-word phrases. Further, in language production humans are also sensitive to the frequency of multi-word phrases. In a production study, Janssen & Barber (2012) found that participants produced frequent multi-word phrases faster than lower frequency phrases, even after taking into account the frequencies of the individual words.

Further, there is also evidence of multi-word storage from the learning literature (Bannard & Matthews, 2008; Siegelman & Arnon, 2015). For example, Siegelman & Arnon (2015) demonstrated that learning is facilitated by attending to the whole utterance, as opposed to attending to each individual word. Specifically, they used an artificial language paradigm to examine adult L2 learners' ability to learn grammatical gender. They found that adults learn grammatical gender better when they are presented with unsegmented utterances rather than segmented utterances. In other words, attending to the entire utterance, rather than learning to compose the utterance word-by-word, facilitated their learning. It seems plausible that if the entire utterance is being attended to, then participants may be learning (i.e., storing) the entire utterance initially. Further, storing larger-than-word chunks may possibly be facilitating the learning of grammatical gender in their study.

### 3.1.2 What Drives Storage?

Despite the evidence of multi-word holistic storage, however, it is still largely unclear what factors drive storage. Humans seem to be sensitive to a variety of statistical information, including both frequency (e.g., Bybee & Scheibman, 1999; Kapatsinski & Radicke, 2009; Lee & Kapatsinski, 2015; Maye & Gerken, 2000) and predictability (e.g, Olejarczuk et al., 2018; Ramscar et al., 2013).

Traditionally, frequency has been assumed to be the driving factor behind multi-word storage. Indeed, most of the examples of storage given so far have been with respect to frequency. Perhaps the most famous series of studies demonstrating this were conducted by Bybee (Bybee, 2003; Bybee & Hopper, 2001; Bybee & Scheibman, 1999). In a series of studies, Bybee and colleagues demonstrated that a variety of words are reduced more in high-frequency contexts than low-frequency contexts (additionally see Kapatsinski, 2021 for further discussion of this). For example, in addition to the earlier examples, *going to* can be reduced in the frequent future marker, *gonna*, but not in the less frequent verb phrase construction describing motion (e.g., **gonna the store*, Bybee, 2003). This mirrors patterns we see on a word-level (which for the most part must be stored). For example, the reduction of vowels to schwa in English is more advanced in high-frequency words than low-frequency words (Bybee, 2003; Hooper, 1976). In other words, for both words and phrases, sound reduction advances more quickly as a function of frequency (i.e., high frequency phrases and high frequency words are both more reduced than their lower frequency counterparts). While this is not surprising for words (which most theories posit have separate representations), it is surprising for phrases which don't necessarily have to be stored holistically.

On the other hand, predictability has not been directly examined much by the Psycholinguistics literature within the context of holistic multi-word storage (c.f. O'Donnell et al., 2009). Additionally, the previous chapter demonstrated that predictability didn't alleviate the effects of local implausibility in reading. To refresh the readers memory, in the previous chapter I examined whether participants were slower to select the first noun in high-predictability compound nouns in locally implausible contexts (i.e., contexts where the first noun in the compound is implausible but where the second noun eliminates the implausibility; see the below sentences) relative to high-predictability compound nouns in locally plausible contexts.

1. **High Predictability Plausible:** Jimmy spread out the peanut butter.
2. **High Predictability Implausible:** Jimmy picked up the peanut butter.

Note that in the implausible condition, the second noun always eliminates the implausibility (i.e., *spread* *out the peanut* is implausible, but *spread out the peanut butter* is not). If high-predictability compound nouns are stored holistically, participants may be able to access the full compound noun upon encountering the first noun, thus overcoming the local implausibility effect (since the second noun in the compound always eliminates the implausibility). The results suggested that the first noun in the compound nouns was read slower in the implausible condition than in the plausible condition. Interestingly, this slowdown was roughly the same regardless of the predictability of the compound noun. That is, there was an increase in reaction time for selecting the first noun in the compound in the implausible condition (relative to the plausible condition) regardless of the predictability of the second noun in the compound noun. Their results suggest that either predictability doesn't drive the holistic storage of compound nouns or that it doesn't facilitate processing in this manner. However they noted that this may be a task effect, since they used the maze task as opposed to an eye-tracking task.

Despite the lack of direct evidence of predictability in the role of multi-word storage, however, predictability has been shown to play a crucial role in learning (Olejarczuk et al., 2018; Ramscar et al., 2013; Saffran et al., 1996). For example, Olejarczuk et al. (2018) demonstrated that when learning new phonetic categories, learners don't just pay attention to co-occurrence rates, but actively try to predict upcoming sounds, suggesting that the learning of phonetic categories is also driven by prediction (i.e., the predictability of a given sound within a context). Further, in learning new words, Ramscar et al. (2013) demonstrated that children are sensitive to how predictable a cue is of an outcome (e.g., a high-frequency cue will be ignored if it isn't predictive of a specific outcome). Additionally, word-segmentation (i.e., learning which segments in an utterance are words) is also highly sensitive to predictability (Saffran et al., 1996). In their classic paper, Saffran et al. (1996) demonstrated that children keep track of transitional probabilities – a measurement of predictability – to segment the speech stream. While these are studies examining learning, not storage, the units that we learn may likely be the units we store. If predictability drives what we learn, it may also drive what we store.

Thus, the current literature presents strong evidence for the role of frequency in the storage of multi-word phrases, as well as suggests the possibility of a further influence of predictability. However, it remains unclear to what extent each of these factors drives storage and whether they interact at all with each other.

### 3.1.3 Representation of Stored Units

Given the evidence that a lot more may be stored than previously thought, another important question to consider is what the internal representations of these units is. Specifically, do the stored units maintain their own internal representation with respect to their component parts? For example, it is possible that the representation of high-frequency phrases, such as *pick up,* retains the representations of the component parts *pick* and *up* (Figure 3.1.3.2). On the other hand, it is possible that the phrase lacks internal representation of the component parts, either because it was lost over time or because it was not learned to begin with.

Indeed, there seems to be some evidence that multi-word phrases may not have a fully intact internal structure with respect to their component parts. For example, Kapatsinski & Radicke (2009) demonstrated that in high frequency V+*up* constructions, it is harder to recognize the segment *up* (with respect to medium-frequency V+*up* constructions). This suggests that those items may have a holistic representation that has lost some of its internal structure. In their study, participants were given different auditory sentences and tasked with pressing a button immediately if they heard the segment *up*. Interestingly, they found that recognizability of *up* follows a U-shaped pattern with respect to the frequency of the phrase. That is, participants were slow to recognize *up* in low frequency phrasal verbs, but for medium-high frequency phrasal verbs they were quicker to recognize *up*. However, upon reaching the highest frequency words participants once again grew slower to recognize *up* (See Figure 3.1.3.1). Though it's important to note that the original paper does not take into account predictability. It's unclear how to account for the increase in recognition time for the highest frequency items if there is no loss of internal representation of those items.

A visualization of what a stored representation with and without internal structure may look like is presented in Figure 3.1.3.2. The left tree represents the phrase *pick up* stored with its internal structure still intact, whereas the right tree represents *pick up* stored without internal structure. Note that both trees are examples of a holistically stored representation. The key difference is whether the internal structure remains intact in the holistic representation. The results from Kapatsinski & Radicke (2009) suggest that for high-frequency verb+*up* collocations, their representation may be more similar to the tree on the right, since participants were slower to recognize *up*. We will revisit this point in the discussion section in more detail.

Figure 3.1.3.1: The U-shaped effect of the frequency of verb+*up* constructions on the speed with which *up* is detected, reproduced from Kapatsinski & Radicke (2009).

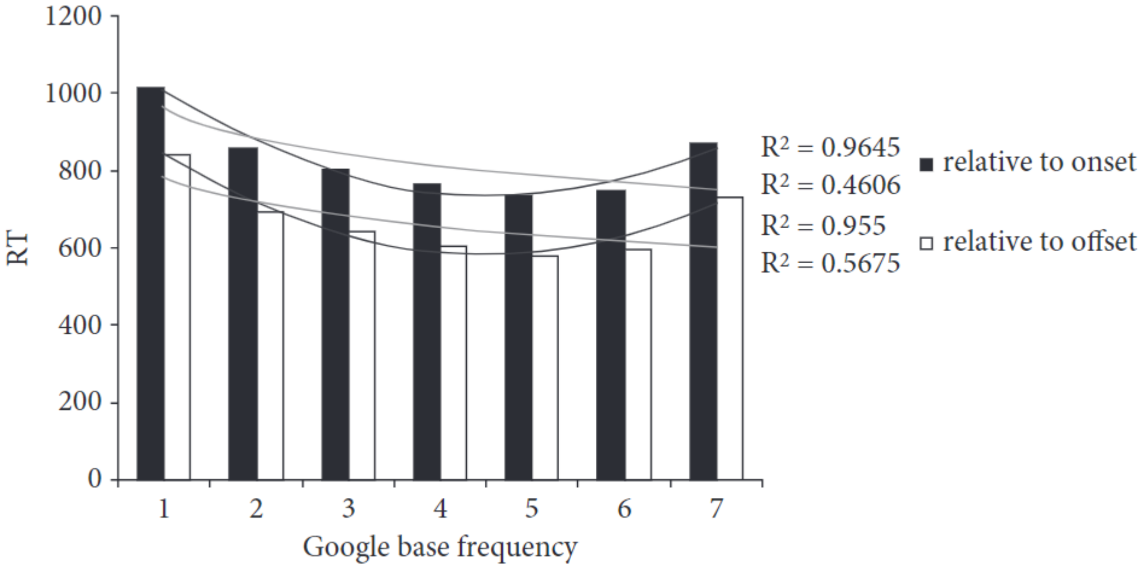
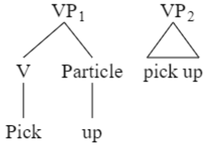


Figure 3.1.3.2: A diagram of two ways the word *pick up* could be stored. The left tree demonstrates a stored representation of *pick up*, where the internal structure is still intact. The right tree demonstrates a holistically stored unit, where there is a loss of internal structure. Note that both of these are stored structures, as opposed to a compositional representation of *pick up* which would be comprised of the individual representations *pick* and *up*.



It's worth noting that in the case of phrasal verbs like *pick up*, it can't be the case that the entire internal representation is lost because it is possible to syntactically alternate it (e.g., *pick up the* *cup* vs *pick the cup up*). However, it is possible that semantic or lemma information is lost in the holistic representation. That is, it is possible that syntactic and/or morphological information may be preserved even if semantic or lemma information is lost. In other words, loss of internal representation may happen at different levels as opposed to being an all-or-nothing process.

### 3.1.4 Present Study

The present study examines the factors that drive storage and the representations of stored items by extending Kapatsinski & Radicke (2009) to look at the effects of both frequency, predictability, and their interaction on the processing of V+*up* phrases. Similar to Kapatsinski & Radicke (2009) , participants are tasked with pressing a button once they hear the segment *up* (which in our study occurs either as a particle within verb phrases, e.g., *pick up*, or part of a word, e.g., *puppet*), but in our case the stimuli varied in frequency, predictability, and whether they were a phrasal verb or not. Since both frequency and predictability effects are rather robust in the literature, we should at the very least see a negative correlation between frequency and predictability and recognition time (up to perhaps a certain point, where recognition time may increase). Further, if predictability is not a driving factor of storage, we should see an increase in recognition times for only the most *frequent* phrases. On the other hand, if predictability does drive storage, we may see an increase in reaction time for both frequent and predictable phrases.

## 3.2 Methods

### 3.2.1 Participants

Participants were recruited through the University of California, Davis Linguistics/Psychology Human Subjects Pool. 350 people participated in this study and were compensated in the form of course credit. All participants self-reported being native English speakers. Additionally, 44 participants were excluded due to an accuracy score below our threshold of 70%, leaving a total of 306 participants for the data analysis.

### 3.2.2 Materials

We searched the Google *n*-grams corpus (Lin et al., 2012) for the most predictable and the highest frequency phrases that matched our criteria of containing a verb immediately followed by the word *up*. We operationalized predictability as the odds ratio of the probability of *up* occurring immediately after the verb to the probability of any other word occurring (Equation 3.1).

$$\frac{\text{count}(\textit{Verb+up})}{\text{count}(\textit{Verb}) - \text{count}(\textit{Verb+up})} \tag{3.1}$$

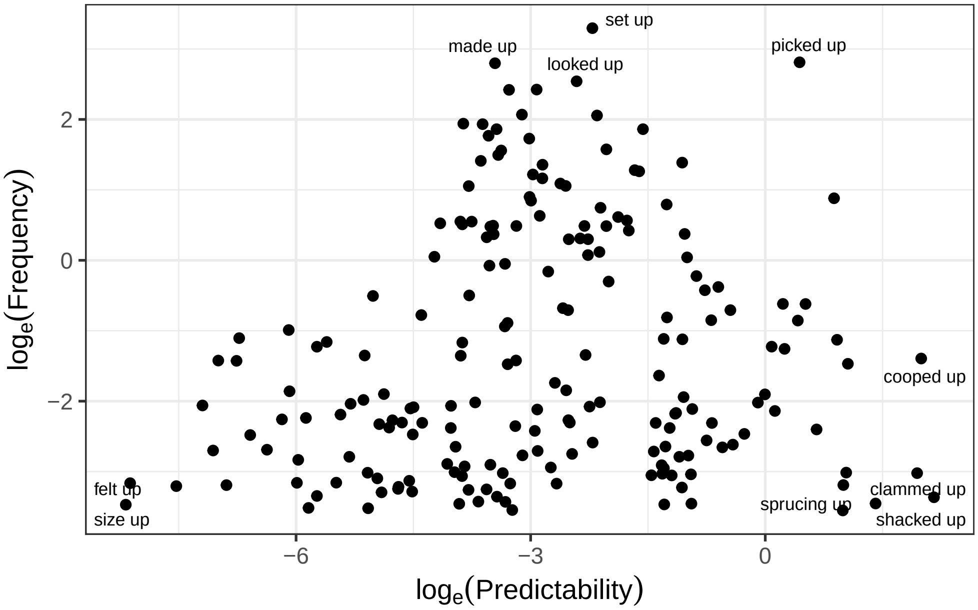
In non-mathematical terms, the above equation quantifies how likely *up* is to follow after the verb relative to every other word that could follow. For example, the odds ratio of *pick up* would be the number of times the entire verb phrase occurs – *pick up* – divided by the number of times the verb – *pick* – occurs without *up* following it.

For the purposes of the present study, we gathered a variety of phrases that varied in both their predictability and frequency and their combination. In order to do this, we extracted the 50 most frequent Verb+*up* items and the 50 most predictable ones. Next, we selected 100 more by randomly sampling from the remaining items. In order to ensure stable predictability estimates we eliminated words that a college-aged speaker wouldn't have heard more than 10 times.[3] We then visually inspected the data to confirm that our data spanned across both the frequency and predictability continuum. This distribution is presented in Figure 3.2.2.1.

Phrasal verbs show a syntactic alternation that is not present in all verb+*up* collocations (e.g., in the example below *lightened up the room* is fine, but *lightened the room up* is weird at best). It is pos-

[3] Levy et al. (2012) extrapolated that the average college-aged speaker has heard about 350 million words in their lifetime. Thus we excluded items that had a frequency smaller than 10 per 350 million.

Figure 3.2.2.1: log-predictability by log-frequency (per million) plot of our items.



sible that due to this syntactic alternation, phrasal verbs may be stored regardless of frequency and predictability. This is because in order to properly use phrasal verbs, a speaker must be aware of the syntactic alternation, which can't simply be predicted compositionally (e.g., some V+*up* phrases are phrasal verbs, while other V+*up* phrases are not phrasal verbs[4]). Thus, we additionally coded our stimuli for whether they were phrasal verbs or not. This coding was done based on whether they could syntactically alternate between having the noun within the verb phrase and having the noun immediately after the verb phrase. For example, since both *pick the cat up* and *pick up the cat* are grammatical, *pick up* was classified as a phrasal verb. Each item was checked by two of the authors. Disagreement was easily resolved by discussion and an agreement was reached for every item.

(1) a. The student lightened up the room.

b. ??The student lightened the room up.

[4] Note that this largely correlates with whether the verb is transitive or not.

We also searched the same corpus for words that contained the segment *up* (e.g., *cupcake*). In order to gather a subset of words that roughly matches the frequency range of our experimental stimuli, we extracted the 50 most frequent words, then sampled from the rest of the dataset to gather an additional 100 words. These 350 items together comprise our stimuli.

For each item, we constructed two sentences: one sentence which contained *up*, and one sentence that was identical except that it didn't include the segment *up*. For words, the entire word was replaced. For phrases, *up* was simply deleted if possible (e.g., *clean up* replaced with *clean*). If this resulted in an awkward sentence, the entire phrase was replaced. An example is given below.

(2) a. He picked up the phone and answered the call.

b. He grabbed the phone and answered the call.

In summary, our stimuli were comprised of 200 Verb+*up* phrases that varied in both frequency and predictability, 150 words that contained *up*, and 350 filler sentences which were matched with our experimental sentences with the exception of having *up* replaced.

After creating the sentences, a native English speaker then recorded each sentence in a random order to minimize any list effect. We subsequently equalized the amplitude such that every sentence was roughly the same loudness.

### 3.2.3 Procedure

Participants were presented with audio sentences via Pavlovia (https://pavlovia.org/), a website for presenting PsychoPy experiments (Peirce et al., 2019). Each participant was presented with 3 practice trials and then 350 sentences. While we had a total of 700 sentences, participants didn't see both the filler and experimental sentence for the same item, thus they only saw half of the stimuli. The order of the sentences was random and exactly half of the sentences contained the target segment (to avoid biasing the participants towards a specific response). Participants were instructed to press a key as soon as they heard the segment *up*, or to press a separate key at the end of the sentence if they did not hear the target segment in the sentence. We then recorded their reaction time of the button press. The experiment took approximately 40 minutes.

## 3.3 Results

The data[5] was analyzed using General Additive Mixed models, as implemented in the *mgcv* package (Wood, 2011) within the R programming environment (R Core Team, 2022). General Additive Mixed Models are models that allow us to model our outcome variable as a combination of the predictors. GAMMs differ from generalized linear regression models in that they allow the predictors to be modeled as non-linear functions, similar to polynomial regression. Specifically, in a Generalized Additive Mixed Model, beta-coefficients are replaced with a smooth function, which is a combination of splines. The more splines that we include, the more wiggly our line will be. In order to avoid overfitting, GAMMs also include a penalty term, $\lambda$, which can be modified to penalize more wiggly lines that aren't justified by the data. While the predictors are allowed to vary non-linearly, the linking function in our case was linear (i.e., response time varied linearly with the spline functions). Our decision to use GAMMs was driven by our hypothesis that recognition times may vary non-linearly as a function of frequency and/or predictability (as suggested by Kapatsinski & Radicke, 2009).

For all of our models, the dependent variable was the time it took for participants to react to the onset of the target segment in experimental sentences/sentences containing *up* (i.e., the time it took participants to press the button after hearing *up*).

In order to visualize the surface of the interaction effect between frequency and predictability, we first ran a model with our independent variable as the interaction between log-predictability

[5] The stimuli, data, and analyses scripts can all be found freely available here: https://github.com/znhoughton/Recognizability-Experiment

and log-frequency, which was allowed to vary non-linearly, and duration of the segment, which was not allowed to vary non-linearly. Additionally, we also included random intercepts for participant, trial, and item, as well as random by-participant slopes for predictability, frequency, their interaction, and trial. All our random-effects were allowed to be wiggly (non-linear). Our model formula is included below in Equation 3.2. This model allows us to visualize the surface of the interaction effect. Note that in GAMMs, the syntax `ti()` is used to model the interaction effects since it produces a tensor product interaction from which the main-effects have been excluded. On the other hand `te()` models the full tensor product smooth without the main-effects excluded. Thus when modeling the main-effects with the interaction effect we use `ti()` and when modeling the surface (that is, without separating the main-effects from the interaction) we use `te()`.

$$\begin{aligned} log(RT) \sim te(Predictability, Frequency) + Duration + s(participant, bs = \text{‘}re\text{’}) + s(Item, bs = \text{‘}re\text{’}) \\ + s(trial, bs = \text{‘}re\text{’}) + s(Predictability, Frequency, participant, bs = \text{‘}re\text{’}) \end{aligned} \tag{3.2}$$

The results of this model are presented in Table 3.1 and visualized in Figure 3.3.0.5. We found no significant effect of the tensor product smooth.[6] Although the tensor product smooth for the interaction effect was not significant, it's possible that phrasal verbs and non-phrasal verbs behave differently and that could be why we don't see an interaction effect. As such, we ran an additional model examining whether the interaction effect was different for phrasal verbs versus non-phrasal verbs. The results for this model are reported in Table 3.2. The results suggest that there is no difference between phrasal and non-phrasal verbs.

It is also possible that despite a lack of an interaction effect, that frequency or predictability independently affect recognition times. Thus, we ran an additional Generalized Additive Model with log-frequency, log-predictability, and the interaction between log-frequency and log-predictability as

[6] We also examined the interaction between frequency and predictability on accuracy (whether they correctly responded to whether *up* was present in the sentence) and similarly found no significant effect.

fixed-effects that could vary non-linearly. Similar to before, duration of the segment was also modeled as a fixed-effect that could not vary non-linearly. The random-effects structure for this model was identical to the previous two models. The model syntax is included below in Equation 3.3:

$$\begin{aligned} log(RT) \sim\ & ti(Predictability) + ti(Frequency) + ti(Predictability, Frequency) + Duration \\ & + s(participant, bs = \text{‘}re\text{’}) + s(Item, bs = \text{‘}re\text{’}) + s(trial, bs = \text{‘}re\text{’}) \\ & + s(Predictability, Frequency, Trial, Participant, bs = \text{‘}re\text{’}) \end{aligned} \tag{3.3}$$

Our results are presented in Table 3.3; Equation 3.3 and visualized in Figure 3.3.0.6. The results demonstrated a significant main-effect of predictability ($p$ < 0.05), but no significant effect of frequency ($p$ = 0.327).[7]

To summarize the results of our generalized additive models, we found no interaction effect between frequency and predictability, no main effect of frequency, but we do find a significant main effect of predictability.

In the Psycholinguistics literature, generalized additive mixed models are not yet well established. Thus, we ran a follow-up Bayesian quadratic regression model to further examine the effects of frequency and predictability on recognition times. Since the Generalized Additive Model suggested that there was no significant interaction between frequency and predictability, we left out the interaction term from the regression model. The random-effects were modeled without correlations between them in order to allow the model to run faster. Equation 3.4 below presents the full model syntax:

[7] We ran a follow-up model without the interaction to determine whether including the interaction effect takes away our power to detect an effect of frequency, however the results for our main-effects are consistent regardless of whether we include the interaction between frequency and predictability in the model.

$$\begin{aligned} log(RT) \sim\ & log(Frequency) + log(Predictability) + Duration + log(Frequency)^2 + log(Predictability)^2 \\ & + (1 + log(Frequency) + log(Predictability) + log(Frequency^2) + log(Predictability^2) \\ & + Duration||Participant) + (1||Item) \end{aligned} \tag{3.4}$$

The results of this model are presented in Table 3.6 and visualized in Figure 3.3.0.8. Following Houghton et al. (2024), in some cases where the credible interval crosses zero, we also report the percentage of posterior samples greater than or less than zero. For the current model, although the credible intervals for both quadratic terms crossed zero, nearly 97% of the posterior samples for predictability$^2$ were greater than zero, and nearly 93% of the posterior samples for frequency$^2$ were greater than zero. A plot of the posterior distribution for each coefficient is presented in Figure 3.3.0.7. The results suggest a U-shaped effect of predictability and a marginal u-shaped effect of frequency on recognition times. In other words, participants recognized *up* faster as frequency or predictability increased, except for the most frequent or most predictable items, where participants were slower to recognize *up*.

Finally, we replicated the analyses from Kapatsinski & Radicke (2009) using two Bayesian quadratic regression models (implemented in *brms;* Bürkner, 2017), one which only included frequency, and one which only included predictability. For the frequency model, the fixed-effects were log-frequency and log-frequency$^2$, along with duration. The model also included random intercepts for participant and item, and random slopes for log-frequency by participant, duration by participant, and log-frequency$^2$ by participant.

The quadratic regression with predictability was identical to the quadratic regression with frequency, except that log-frequency was replaced with log-predictability, and log-frequency$^2$ was replaced with log-predictability$^2$. The random-effects were modeled without correlations between them for both models (this was done to allow the model to run faster, since we collected a large amount of data).

The model syntax for both models is included below in Equation 3.5 and Equation 3.6:

$$\begin{aligned}log(RT) &\sim log(Frequency) + Duration + log(Frequency)^2 \\ &+ (1 + log(Frequency) + log(Frequency)^2 + Duration||Participant) + (1||Item)\end{aligned} \tag{3.5}$$

$$\begin{aligned}log(RT) &\sim log(Predictability) + Duration + log(Predictability)^2 \\ &+ (1 + log(Predictability) + log(Predictability)^2 + Duration||Participant) + (1||Item)\end{aligned} \tag{3.6}$$

The results of our first model are presented in Table 3.4. While the credible interval for log(frequency)$^2$ crosses zero, over 95% of the posterior samples were greater than zero, suggesting an effect of frequency$^2$ on recognition times. A visualization of the model predictions is presented in Figure 3.3.0.3 and a visualization of the posterior distribution is presented in Figure 3.3.0.1.

The results of our second model are presented in Table 3.5. While the credible interval for log(predictability)$^2$ crosses zero, over 96% of the posterior samples were greater than zero, suggesting a meaningful effect. A visualization of the model predictions is included in Figure 3.3.0.4 and a visualization of the posterior distribution is presented in Figure 3.3.0.2.

In summary, our results suggest that when considered independently, there appears to be a U-shaped effect for both frequency and predictability. The effect for frequency is not as reliably detected when predictability is also accounted for in our models, however we do find weak evidence for it. We do not find strong evidence for an interaction between frequency and predictability but it is possible that our study simply does not have the power to detect an interaction effect.

Table 3.1: Model results for the generalized Additive Mixed Model cotanining only the interaction between frequency and predictability.

| | edf | Ref.df | F | p-value |
|---|---|---|---|---|
| te(log-predictability, log-frequency) | 5.59 | 5.73 | 1.86 | 0.090 |
| s(trial) | 0.99 | 1.00 | 115.38 | <0.001 |
| s(participant) | 296.00 | 305.00 | 39.74 | <0.001 |
| s(item) | 175.44 | 195.00 | 10.68 | <0.001 |
| s(log-predictability, log-frequency, trial, participant) | 43.00 | 306.00 | 0.46 | 0.100 |

Table 3.2: Model results for the Generalized Additive Mixed Model cotaining the interaction between frequency and predictability for phrasal vs nonphrasal verbs.

| | edf | Ref.df | F | p-value |
|---|---|---|---|---|
| te(log-predictability, log-frequency):Nonphrasal | 3.93 | 3.98 | 1.46 | 0.210 |
| te(log-predictability, log-frequency):Phrasal | 4.07 | 4.12 | 1.27 | 0.240 |
| s(trial) | 0.99 | 1.00 | 115.65 | <0.001 |
| s(participant) | 295.99 | 305.00 | 39.83 | <0.001 |
| s(item) | 172.59 | 191.00 | 10.94 | <0.001 |
| s(log-predictability, log-frequency, trial, participant) | 42.97 | 306.00 | 0.46 | 0.100 |

Table 3.3: Model results for the Generalized Additive Mixed Model cotaining Frequency, Predictability, and the interaction between them.

| | edf | Ref.df | F | p-value |
|---|---|---|---|---|
| ti(log-frequency) | 2.16 | 2.20 | 1.73 | 0.270 |
| ti(log-predictability) | 1.97 | 2.01 | 4.10 | 0.020 |
| ti(log-frequency, log-predictability) | 1.00 | 1.00 | 0.89 | 0.350 |
| s(participant) | 296.33 | 305.00 | 37.72 | <0.001 |
| s(item) | 175.70 | 195.00 | 10.76 | <0.001 |
| s(log-predictability, log-frequency, participant) | 0.17 | 305.00 | 0.00 | 0.600 |

Table 3.4: Results for the Bayesian quadratic regression model containing only frequency and frequency$^2$.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | -0.102 | 0.025 | -0.150 | -0.054 | 0.000 |
| log-frequency | 0.016 | 0.011 | -0.005 | 0.038 | 93.310 |
| Duration | -0.084 | 0.098 | -0.274 | 0.108 | 19.355 |
| log-frequency^2 | 0.006 | 0.004 | -0.001 | 0.013 | 95.225 |

Table 3.5: Results for the Bayesian quadratic regression model containing only predidctability and predictability$^2$.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | -0.110 | 0.027 | -0.163 | -0.058 | 0.0000 |
| log-predictability | 0.008 | 0.011 | -0.014 | 0.029 | 75.7350 |
| Duration | -0.089 | 0.098 | -0.280 | 0.102 | 18.4225 |
| log-predictability^2 | 0.003 | 0.002 | -0.000 | 0.006 | 96.0975 |

Table 3.6: Model results for the Bayesian quadratic regression model containing fixed-effects for frequency, predictability, and their quadratics.

| | Estimate | Est.Error | Q2.5 | Q97.5 | % Samples > 0 |
|---|---|---|---|---|---|
| Intercept | -0.102 | 0.029 | -0.161 | -0.046 | 0.02625 |
| log-frequency | 0.019 | 0.011 | -0.002 | 0.041 | 96.15625 |
| log-predictability | 0.009 | 0.011 | -0.013 | 0.032 | 78.99750 |
| duration | -0.135 | 0.098 | -0.328 | 0.057 | 8.27125 |
| log-predictability^2 | 0.003 | 0.002 | -0.000 | 0.007 | 96.88125 |
| log-frequency^2 | 0.005 | 0.004 | -0.002 | 0.012 | 92.94375 |

Figure 3.3.0.1: Plot of the posterior distribution for the beta value of each fixed-effect in our frequency-only quadratic regression model. The y-axis contains the different fixed-effects and the x-axis contains the posterior distribution of beta values for the corresponding fixed-effect.

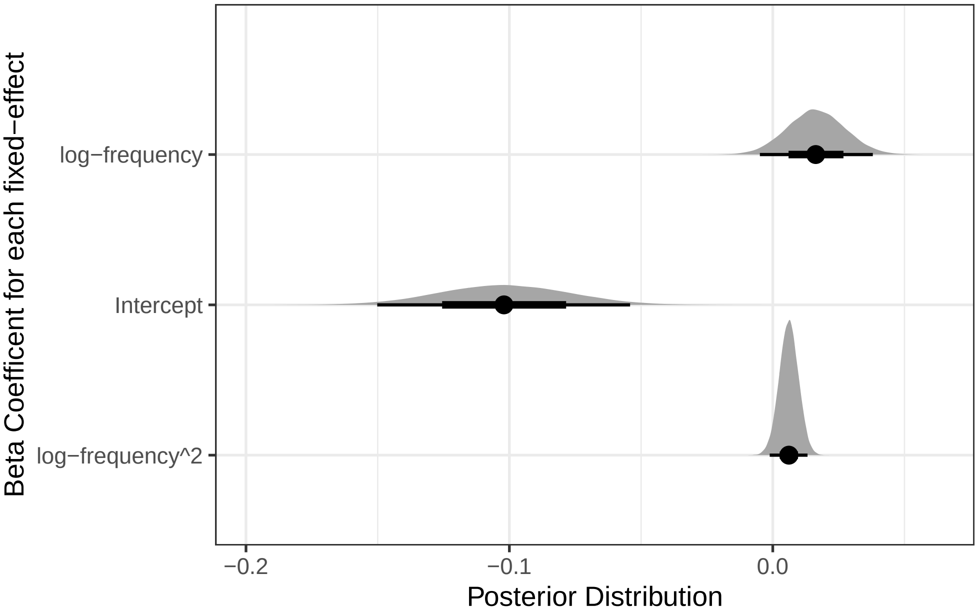


Figure 3.3.0.2: Plot of the posterior distribution for the beta value of each fixed-effect in our predictability-only quadratic regression model. The y-axis contains the different fixed-effects and the x-axis contains the posterior distribution of beta values for the corresponding fixed-effect.

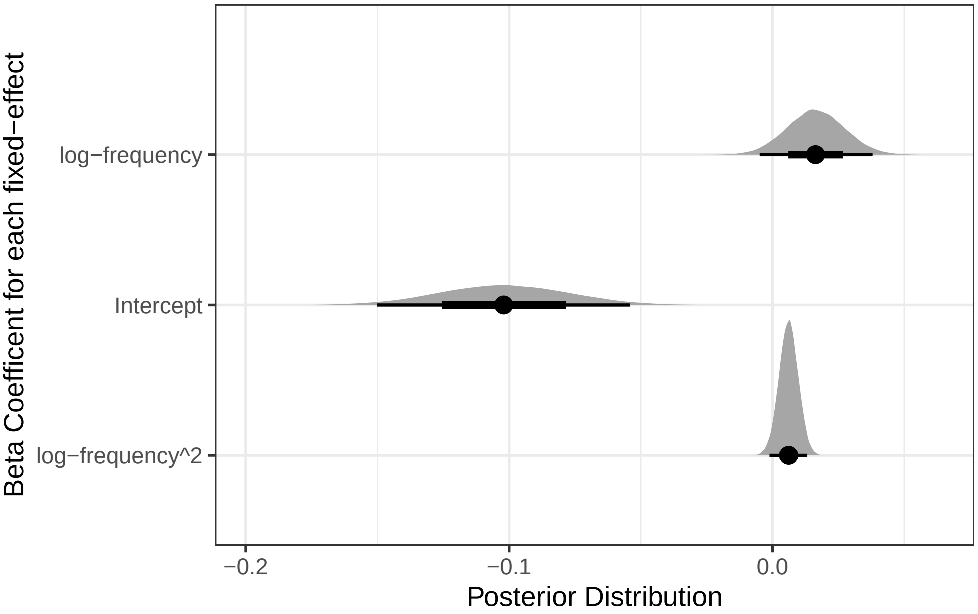


Figure 3.3.0.3: Model predictions for the effects of frequency on reaction times for the frequency-only Bayesian quadratic model.

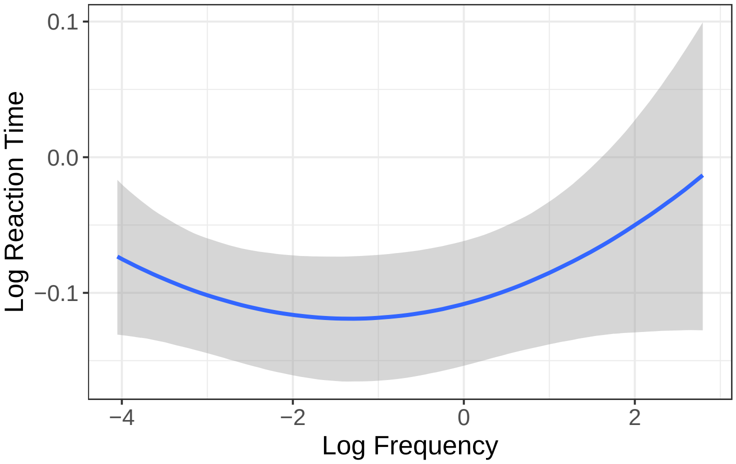


Figure 3.3.0.4: Model predictions for the effect of predictability on reaction times for the predictability-only models.

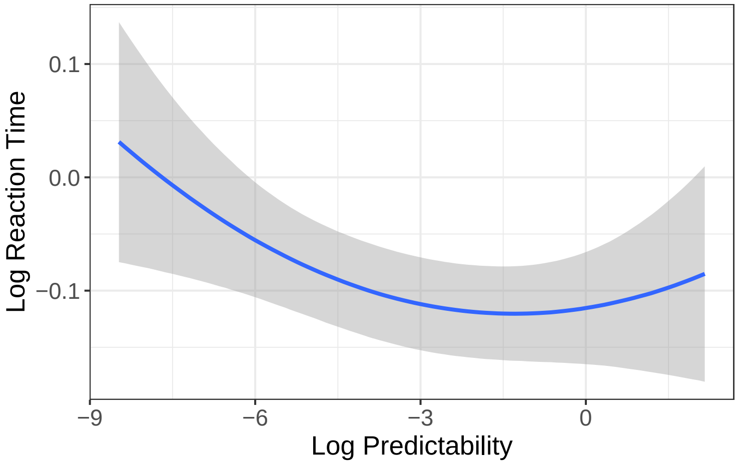


Figure 3.3.0.5: Plot of the interaction effect between predictability and frequency of our GAM model containing only the interaction between frequency and predictability. The brightness of the coloration denotes the strength of the effect at the point in the graph. Brighter colors denote longer reaction times.

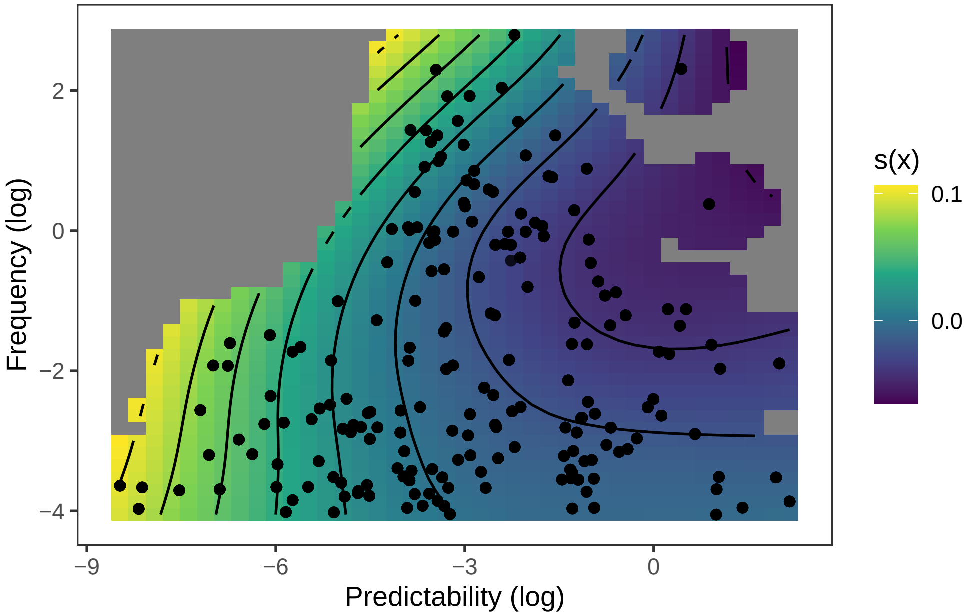


Figure 3.3.0.6: Plot of our GAM model's predicted effect of log(predictability) on recognition time.

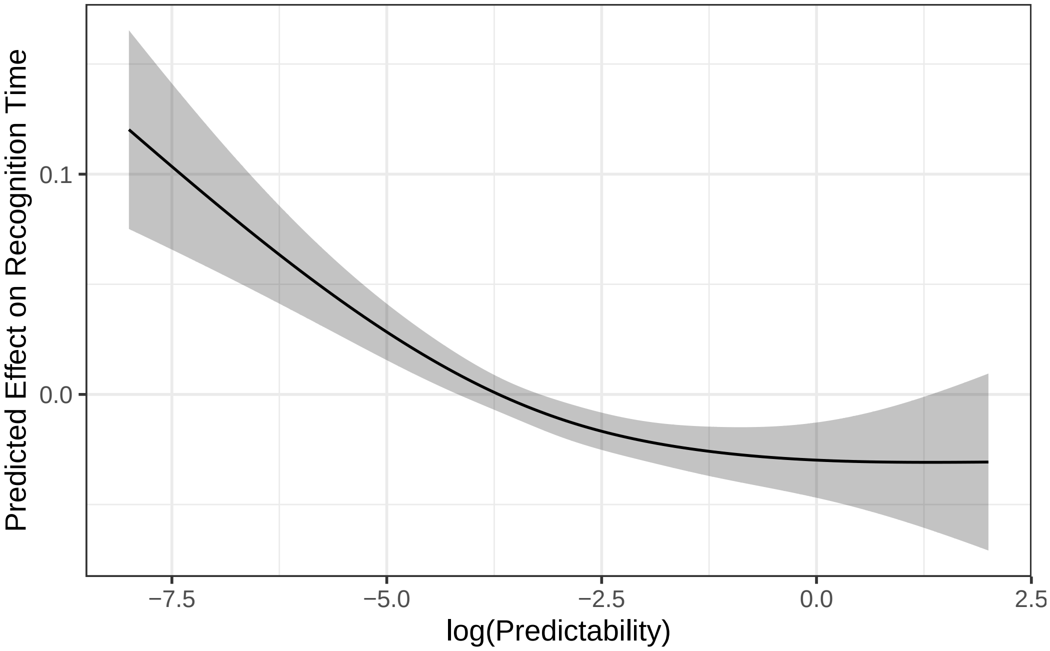


Figure 3.3.0.7: Plot of the posterior distribution for the beta value of each fixed-effect in our Bayesian quadratic regression model. The y-axis contains the different fixed-effects and the x-axis contains the posterior distribution of beta values for the corresponding fixed-effect.

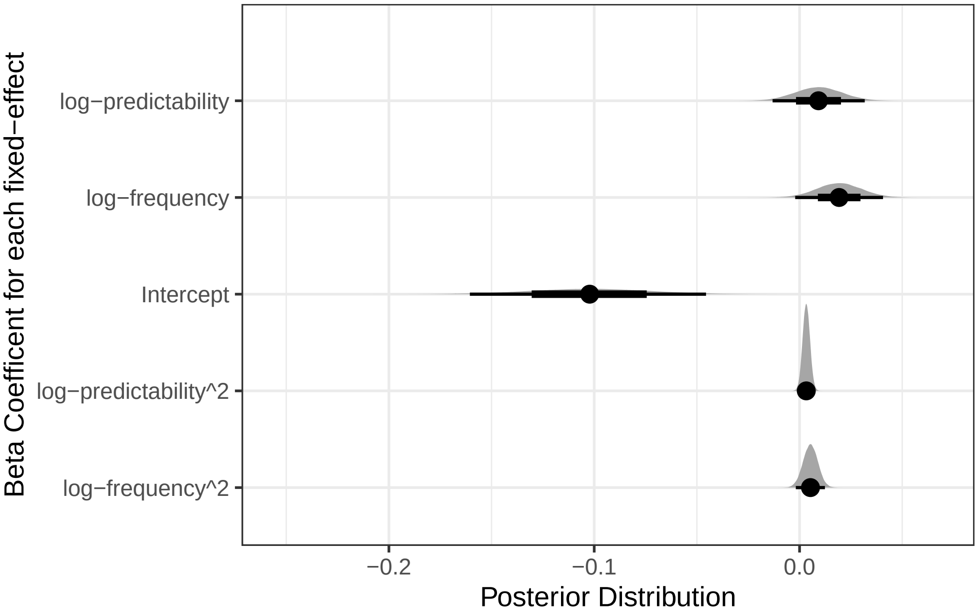
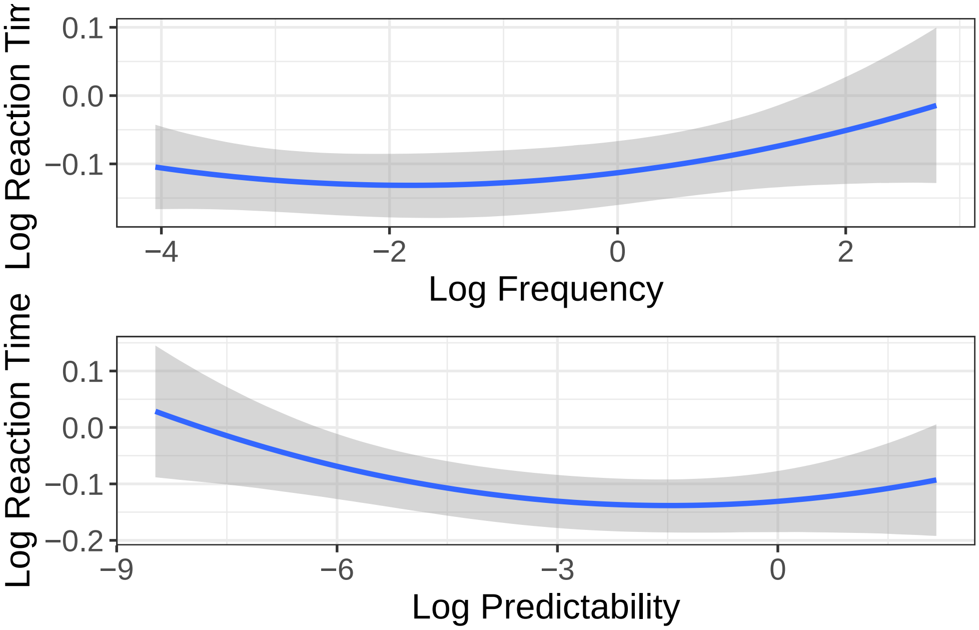


Figure 3.3.0.8: Visualization of the model results from Table 3.6 for frequency (top) and predictability (bottom). Frequencies are per million.



## 3.4 Discussion

The present study examined the effects of frequency and predictability on the recognizability of the particle *up* in English phrasal verbs. We found a U-shaped effect for both frequency and predictability on recognizability: as frequency and predictability increased, people were faster at recognizing *up*, until reaching the highest frequency/most predictable items, where people were slower. Additionally, we also found no meaningful differences between phrasal verbs (e.g., *pick up*) and non-phrasal verbs (e.g., *stir up*), suggesting that this slowdown is due to statistical properties of the language as opposed to syntactic properties.

There are three possible accounts for the slowdown we see for the highest frequency or predictability items. First, it's possible that people are attending less to *up* or even skipping it in high frequency and high predictability phrases. This account, unlike the other accounts that we'll discuss, does not explicitly require the high frequency and high predictability phrases to be stored. Instead, the listener may be able to process the meaning of the phrase fast enough that they don't need to wait to hear the entire phrase. For example, it's possible that for high-frequency and high-predictability items, when accessing the first word, e.g., *pick*, the listener accesses the representation of the entire phrase — either a holistic representation or a compositional representation — immediately, before even hearing *up*. The listener can then continue to process the next words (skipping over *up*). Since the task is to respond when they hear *up*, the delay in reaction time may be because they're not accessing the phonological representation of *up*. Instead, they may access the semantic representation of the phrase without initially accessing the phonological representation of *up* and go on to recover the phonological representation from the semantic representation of the phrase, causing a delay in recognition time. Indeed, this possibility was suggested by Healy (1976), who suggested that in reading once people process the meaning of a word, they move on to the next word regardless of whether they have processed each individual letter. This account doesn't explicitly require *pick up* to be stored holistically since a listener could hear *pick*, predict *up*, and compose the meaning *pick up* despite having not heard *up*. However, it also isn't incompatible with a storage account, since the listener might hear *pick,* predict *up*, and then accesses a stored holistic representation of *pick up*. In other words, if listeners are attending less to *up*, then it's unclear whether the listeners are accessing a representation formed by a compositional process (i.e., accessing *pick,* predicting *up,* and composing *pick up*) or simply retrieving a stored form from memory (accessing a holistic representation *pick up*).

The next two accounts all require the high-frequency and high-predictability items to be stored holistically, but vary with respect to whether the holistically stored representations retain their internal structure.

It is possible that the slowdown for the high frequency and high predictability items is due to competition between an additional representation. This competition can either be between a holistic representation that has internal structure and a compositional representation, or between a holistic representation that does not have internal structure and the compositional representation. Compositional representation here refers to a representation that is formed by accessing individual forms (e.g., *pick* and *up*) and combining them via some generative process. High-frequency and high-predictability items may develop a holistic representation separate from the compositional representation and this additional representation may compete with the compositional representation causing the slowdown. This account doesn't necessarily need to involve a loss of internal structure because simply having an additional representation to compete with can result in a slowdown, however it also not incompatible with an account where the holistic representation has lost some of its internal structure. These two possibilities both account for the slowdown at the highest frequency and highest predictability items.

To break it down further, there is a good deal of evidence that different mental representations compete for recognition (Oppenheim & Balatsou, 2019; c.f. Staub et al., 2015). A representation is selected once it receives sufficiently more activation than its competitors (McClelland & Rumelhart, 1981). For example, in picture-naming tasks in which participants are tasked with naming a picture while confronted with a distractor word, participants are generally slower to produce the intended word when the distractor word is semantically related to the picture (McClelland & Rumelhart, 1981; Schriefers et al., 1990; Starreveld & La Heij, 1995). This effect is not restricted to production as we see similar competition effects in comprehension as well. For example, Magnuson et al. (2007) examined the role of competition in word recognition using a visual world paradigm, where participants saw words on a screen and were instructed to select the word that they heard. To measure word-recognition, an eye-tracker was used to track pupil fixations. In each of the trials there was a single distractor image. They found that words with low cohort density (i.e., words that have fewer phonological competitors) showed a larger proportion of target to nontarget fixations. That is, participants looked the distractor image less relative to the target word when the word had fewer competitors. Given the inhibitory effects of competition, it is possible that the delay in reaction time for *up* in high-frequency and predictability phrases may be a consequence of an additional representation competing with the compositional representation. However, there is also evidence that competition has no effect on comprehension (Staub et al., 2015). Using reaction time data from a cloze completion task, Staub et al. (2015) demonstrated that a RACE model with neither facilitation nor inhibition between competitors can account for the data. Thus the evidence for competition effects in comprehension is mixed. Note that this account is agnostic about whether the holistic representation has lost its internal structure or not: simply having an additional representation to compete with can cause the slowdown.

Lastly it is possible that rather than being driven by competition, listeners are simply accessing a holistically stored representation of the phrase that lacks internal structure. This interpretation seems quite likely given that we see a U-shaped effect in both phrasal (e.g., *pick up*) and non-phrasal verbs (e.g., *stir up*). Phrasal verbs have a syntactic alternation that may lead to all of them being stored, regardless of whether they are frequent/predictable or not. For example, In a corpus study, Hampe (2012) argued that *Verb-Object-Particle* constructions and *Verb-Particle-Object* constructions are two distinct constructions.[8] If the increase in reaction time is simply due to competition between the holistically stored representation and the individual word-level representations, then if all phrasal verbs are stored we would expect all of the phrasal verbs to be recognized more slowly. This is because all of the phrasal verbs, regardless of frequency, would have an additional representation that would compete for activation. However, we only see a slowdown for the most frequent or most predictable phrases, suggesting that storage alone isn't driving the effect. Instead, it is the combination of storage and usage that leads to loss of internal representation.

One explanation for why high frequency and high predictability items may not have an intact internal representation is that the internal structure for those items may never have been learned to begin with. Children are experts at statistical learning and use transitional probabilities to divide the continuous speech stream (Saffran et al., 1996). High predictability phrases in the present study, by definition, have higher transitional probabilities between words. Thus if children are relying on transitional probabilities to separate speech into individual words, the individual words in the most predictable phrases may not be separated out of the speech stream initially.

Further, many high-frequency (e.g., *set up*) and high-predictability (e.g., *conjure up*) phrases have semantically vague relationships that might make it difficult to split them up on a semantic basis. It seems plausible then that maybe these phrases weren't learned as being composed of individual words initially and thus the internal structure for the holistically stored items may not have been learned.

[8] However, the same study also makes the claim that these templates are different from more lexically specific constructions, thus it is unclear in what ways these templates may pattern similarly to holistically stored lexical items.

The example, *trick or treat*, is a prime example of a phrase that does not seem to have a clear semantic relationship between the phrase and its component parts.

On the other hand, the internal structure may have been lost over time. For example, Harmon & Kapatsinski (2017) demonstrated that as learners repeatedly experience a form with a specific meaning, they become more likely to use that form to express novel meanings in production (resulting in semantic extension). It is possible that this accessibility effect similarly drives a loss of internal structure: As a phrase becomes more semantically extended, the internal structure may be lost over time. That is, as a phrase such as *pick up* becomes extended to express novel meanings such as *continue* ("Let's pick up from where we last left off"), the relationship between the phrase and its internal pieces (e.g., the relationship between *pick up* and the individual words *pick* and *up*) becomes less transparent, and the learner may slowly unlearn this relationship as it becomes less useful.

In summary, our results suggest that both frequency and predictability may drive the holistic storage of phrasal verbs, and these holistically stored items may compete with their component parts during lexical access. However, future work is still needed to confirm whether the slowdown for the highest frequency and highest predictability items is indeed due to a stored holistic representation or if it's due to shallower attention mechanisms.

# References

Al-Selwi, S. M., Hassan, M. F., Abdulkadir, S. J., & Muneer, A. (2023). LSTM inefficiency in long-term dependencies regression problems. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *30*(3), 1631. https://www.researchgate.net/profile/Amgad-Muneer-2/publication/370769893_LSTM_Inefficiency_in_Long-Term_Dependencies_Regression_Problems/links/64624134f43b8a29ba525b9b/LSTM-Inefficiency-in-Long-Term-Dependencies-Regression-Problems.pdf

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5-6), 509–559. https://doi.org/10.1177/0142723719869731

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Baayen, H., Schreuder, R., De Jong, N., & Krott, A. (2002). *Dutch inflection: The rules that prove the exception* (S. Nooteboom, F. Weerman, & F. Wijnen, Eds.; Vol. 30, pp. 61–92). Springer Netherlands. http://link.springer.com/10.1007/978-94-010-0355-1_3

Baayen, R. H., Milin, P., rević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481. https://doi.org/10.1037/a0023851

Bannard, C., & Matthews, D. (2008). Stored Word Sequences in Language Learning: The Effect of Familiarity on Children's Repetition of Four-Word Combinations. *Psychological Science*, *19*(3), 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

https://doi.org/10.1016/j.jml.2012.11.001

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623. https://doi.org/10.1145/3442188.3445922

Bengio, Y., Frasconi, P., & Simard, P. (1993). *IEEE international conference on neural networks*. 1183–1188 vol.3. https://doi.org/10.1109/ICNN.1993.298725

Benor, S. B., & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of english binomials. *Language*, *82*(2), 233–278. https://doi.org/10.1353/lan.2006.0077

Berko, J. (1958). The Child's Learning of English Morphology. WORD, *14*(2-3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*(November 2019), 104082. https://doi.org/10.1016/j.jml.2019.104082

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 128. https://www.jstatsoft.org/article/view/v080i01

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*(3), 261–290. https://doi.org/10.1017/S0954394502143018

Bybee, J. (2003). *Phonology and language use* (Vol. 94). Cambridge University Press.

Bybee, J., & Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. *Typological Studies in Language*, *45*, 126. https://www.torrossa.com/gs/resourceProxy?an=5002168&publisher=FZ4850#page=10

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of don't in english. *Linguistics*, *37*(4). https://doi.org/10.1515/ling.37.4.575

Chomsky, N. (1965). *Aspects of the theory of syntax special technical report no. 11*.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*(3), 542–551.

https://doi.org/10.1111/tops.12274

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences, 7*(5), 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., et al. (2024). Olmo: Accelerating the science of language models. *arXiv Preprint arXiv:2402.00838.*

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (n.d.). *Colorless green recurrent networks dream hierarchically*. https://doi.org/10.48550/arXiv.1803.11138

Haley, C. (2020). *This is a BERT. Now there are several of them. Can they generalize to novel words?* 333341. https://aclanthology.org/2020.blackboxnlp-1.31/

Hampe, B. (2012). Transitive phrasal verbs in acquisition and use: A view from construction grammar. *Language Value, 4*(1), 1–32. https://raco.cat/index.php/LanguageValue/article/view/302086

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology, 98,* 22–44. https://doi.org/10.1016/j.cogpsych.2017.08.002

Healy, A. F. (1976). Detection errors on the word the: Evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception and Performance, 2*(2), 235. https://psycnet.apa.org/journals/xhp/2/2/235/

Hooper, J. B. (1976). Word frequency in lexical diffusion and the source of morphophonological change. *Current Progress in Historical Linguistics, 96,* 105.

Houghton, Z., Kato, M., Baese-Berk, M., & Vaughn, C. (2024). Task-dependent consequences of disfluency in perception of native and non-native speech. *Applied Psycholinguistics,* 1–17. https://doi.org/10.1017/S0142716423000486

Janssen, N., & Barber, H. A. (2012). Phrase Frequency Effects in Language Production. *PLOS ONE, 7*(3), e33202. https://doi.org/10.1371/journal.pone.0033202

Kapatsinski, V. (2018). *Changing minds changing tools: From learning theory to language acquisition to language change*. MIT Press.

Kapatsinski, V. (2021). Hierarchical inference in sound change: Words, sounds, and frequency of use. *Frontiers in Psychology*, *12*(August). https://doi.org/10.3389/fpsyg.2021.652664

Kapatsinski, V., & Radicke, J. (2009). *Frequency and the emergence of prefabs: Evidence from monitoring*. *January 2009*, 499. https://doi.org/10.1075/tsl.83.14kap

Lasri, K., Seminck, O., Lenci, A., & Poibeau, T. (2022). Subject verb agreement error patterns in meaningless sentences: Humans vs. BERT. *arXiv Preprint arXiv:2209.10538*.

Lee, O., & Kapatsinski, V. (2015). *Frequency effects in morphologisation of korean /n/-epenthesis*. 1–23.

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in english. *Cognition*, *122*(1), 12–36. https://doi.org/10.1016/j.cognition.2011.07.012

Li, B., & Wisniewski, G. (2021). *Are neural networks extracting linguistic properties or memorizing training data? An observation with a multilingual probe for predicting tense*. https://shs.hal.science/halshs-03197072/

Li, B., Wisniewski, G., & Crabbé, B. (2023). Assessing the capacity of transformer to abstract syntactic representations: A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, *11*, 18–33. https://doi.org/10.1162/tacl_a_00531

Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., Brockman, W., & Petrov, S. (2012). *Syntactic annotations for the google books ngram corpus*. 169174. https://aclanthology.org/P12-3029.pdf

Liu, Z., & Morgan, E. (2020). *Frequency-dependent regularization in constituent ordering preferences.* https://www.cognitivesciencesociety.org/cogsci20/papers/0751/0751.pdf

Liu, Z., & Morgan, E. (2021). *Frequency-dependent regularization in syntactic constructions*. 387389. https://aclanthology.org/2021.scil-1.41.pdf

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The Dynamics of Lexical Competition During Spoken Word Recognition. *Cognitive Science*, *31*(1), 133–156. https://doi.org/10.1080/03640210709336987

Maye, J., & Gerken, L. (2000). *Learning phonemes without minimal pairs*. *2*, 522533. https://www.academia.edu/download/68237640/Learning_Phonemes_Without_Minimal_Pairs20210721-21044-1t0fvya.pdf

McClelland, J. L., Elman, J. L., & LANGUAGE, C. U. S. D. L. J. C. F. R. I. (1984). The TRACE model of speech perception. *California University San Diego, La Jolla Center for Research in Language*. https://apps.dtic.mil/sti/citations/ADA157550

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375. https://psycnet.apa.org/record/1981-31825-001

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? Evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics, 11,* 652670. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00567/116616

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language, 60*(1), 65–91. https://doi.org/10.1016/j.jml.2008.07.002

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science, 331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Misra, K., & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. *arXiv Preprint arXiv:2403.19827*.

Mollica, F., & Piantadosi, S. T. (2019). Humans store about 1.5 megabytes of information during language acquisition. *Royal Society Open Science, 6*(3), 181393. https://doi.org/10.1098/rsos.181393

Morgan, E., & Levy, R. (2015a). *Modeling idiosyncratic preferences : How generative knowledge and expression frequency jointly determine language structure*. 1649–1654.

Morgan, E., & Levy, R. (2015b). *Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure.*

Morgan, E., & Levy, R. (2016a). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition, 157*, 384–402. https://doi.org/10.1016/j.cognition.2016.09.011

Morgan, E., & Levy, R. (2016b). Frequency-dependent regularization in iterated learning. *The Evolution of Language: Proceedings of the 11th International Conference*.

Morgan, E., & Levy, R. (2024). Productive knowledge and item-specific knowledge trade off as a function of frequency in multiword expression processing. *Language, 100*(4), e195–e224. https://muse.jhu.edu/pub/24/article/947046

Nooteboom, S., Nooteboom, S. G., Weerman, F., & Wijnen, F. N. K. (2002). *Storage and computation in the language faculty*. Springer Science & Business Media. https://books.google.com/books?hl=en&lr=&id=Sa_dGP0AT-YC&oi=fnd&pg=PR7&dq=nootebom+storage+and+computation&ots=nYGC8JjkTW&sig=1RZzWemOQIzn6bmSH7NF396lKZQ

O'Donnell, T. J. (2016). Productivity and reuse in language. *Productivity and Reuse in Language*, 1613–1618. https://doi.org/10.7551/mitpress/9780262028844.001.0001

O'Donnell, T. J., Tenenbaum, J. B., & Goodman, N. D. (2009). *Fragment Grammars: Exploring Computation and Reuse in Language*. https://dspace.mit.edu/handle/1721.1/44963

Olejarczuk, P., Kapatsinski, V., & Baayen, R. H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard, 4*(s2), 1–9. https://doi.org/10.1515/lingvan-2017-0020

Oppenheim, G. M., & Balatsou, E. (2019). Lexical competition on demand. *Cognitive Neuropsychology, 36*(5-6), 216–219. https://doi.org/10.1080/02643294.2019.1580189

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*, 195–203.

Pierrehumbert, J. B. (2001). *Exemplar dynamics: Word frequency, lenition and contrast* (J. L. Bybee & P. J. Hopper, Eds.; Vol. 45, p. 137). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.45.08pie

Pierrehumbert, J. B. (2016). Phonological Representation: Beyond Abstract Versus Episodic. *Annual Review of Linguistics, 2*(Volume 2, 2016), 33–52.

https://doi.org/10.1146/annurev-linguistics-030514-125050

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*(11), 456–463. https://doi.org/10.1016/S1364-6613(02)01990-3

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017–1023. https://doi.org/10.1177/0956797612460691

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102. https://doi.org/10.1016/0749-596X(90)90011-N

Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60–75. https://doi.org/10.1016/j.jml.2015.07.003

Siyanova-Chanturia, A., Conklin, K., & Heuven, W. J. B. van. (2011). Seeing a phrase " time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *37*(3), 776–784. https://doi.org/10.1037/a0022531

Smith, N. (2014). ZS: A file format for efficiently distributing, using, and archiving record-oriented datasets of any size. *Vorpus.Org*, *270273*(270273), 1–39.

Starreveld, P. A., & La Heij, W. (1995). Semantic interference, orthographic facilitation, and their interaction in naming tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 686. https://psycnet.apa.org/record/1995-42762-001

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004

Staub, A., Rayner, K., Pollatsek, A., Hyönä, J., & Majewski, H. (2007). The time course of plausibility effects on eye movements in reading: Evidence from noun-noun compounds. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(6), 1162–1169. https://doi.org/10.1037/0278-7393.33.6.1162

Stemberger, J. P., & MacWhinney, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition*, *14*(1), 17–26. https://doi.org/10.3758/BF03209225

Stemberger, J. P., & MacWhinney, B. (2004). Are inflected forms stored in the lexicon. *Morphology: Critical Concepts in Linguistics*, *6*, 107122.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv Preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/7181-attention-is-all

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, *60*(3), 158189. https://www.sciencedirect.com/science/article/pii/S0010028509000826?casa_token=gRo7ST2VYTQAAAAA:EgIkNTP-CucMTpDL6ttAuK08KLqCfi9cmrdx3sTTpZiLjrC6T5qMm1vM0girWG6lHDl9Vjpk

Wang, Y., Liu, D., & Wang, Y. (2003). Discovering the Capacity of Human Memory. *Brain and Mind*, *4*(2), 189–198. https://doi.org/10.1023/A:1025405628479

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). *Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora* (A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell, Eds.; p. 134).

Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-babylm.1

Weissweiler, L., Mahowald, K., & Goldberg, A. (2025). Linguistic generalizations are not rules: Impacts on evaluation of LMs. *arXiv Preprint arXiv:2502.13195*.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(1), 3–36.

Yao, Q., Misra, K., Weissweiler, L., & Mahowald, K. (2025). Both direct and indirect evidence contribute to dative alternation preferences in language models. *arXiv Preprint arXiv:2503.20850*.

Yi, B. W. (2002). 음운 현상과 빈도 효과.

Zang, C., Wang, S., Bai, X., Yan, G., & Liversedge, S. P. (2024). Parafoveal processing of chinese four-character idioms and phrases in reading: Evidence for multi-constituent unit hypothesis. *Journal of Memory and Language*, *136*, 104508. https://doi.org/10.1016/j.jml.2024.104508

Zwitserlood, P. (2018). *Processing and representation of morphological complexity in native language comprehension and production*. 583–602. https://doi.org/10.1007/978-3-319-74394-3_20