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

2025

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 김선생님 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—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, 2015, 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 multi-morphemic words and multi-word phrases, is stored holistically (e.g., Bybee, 2003; Morgan & Levy, 2015, 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 that 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, the word *cat* would be stored and *cats* would be generated using knowledge of the grammar (and would not be stored holistically). Similarly, multi-word 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 thought, 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 not the limiting factor that we once thought.

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, 2015, 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 Dr. Janet Pierrehumbert and Dr. 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 the contexts that they have been experienced in. 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. Instead, humans must have representations that enable this sensitivity to within-category variation.

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 multimorphemic 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 a great deal of evidence for 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 (2015) 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 (2015) 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, in a follow-up study Morgan & Levy (2024) demonstrated that these generative preferences exert an effect on all items, even high-frequency items (although generative preferences exert a weaker effect on the high-frequency items than the low-frequency ones).

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). 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 entirely predict the ordering preferences for high-frequency items. Thus, if people are simply composing binomials on the fly from the 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). 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 for holistic storage, it is still unclear what drives storage. Frequency has been assumed – oftentimes implicitly – 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, 2015, 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, in addition to frequency, 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, demonstrate the importance of predictability in learning and 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 3.1.1). 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.

A visualization of what this may look like is demonstrated in Figure 1.4.2. 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.

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) and it is unclear how this would not lead to holistic storage of these items. For

Figure 1.4.1.: A plot of the results reproduced from Kapatsinski & Radicke (2009).

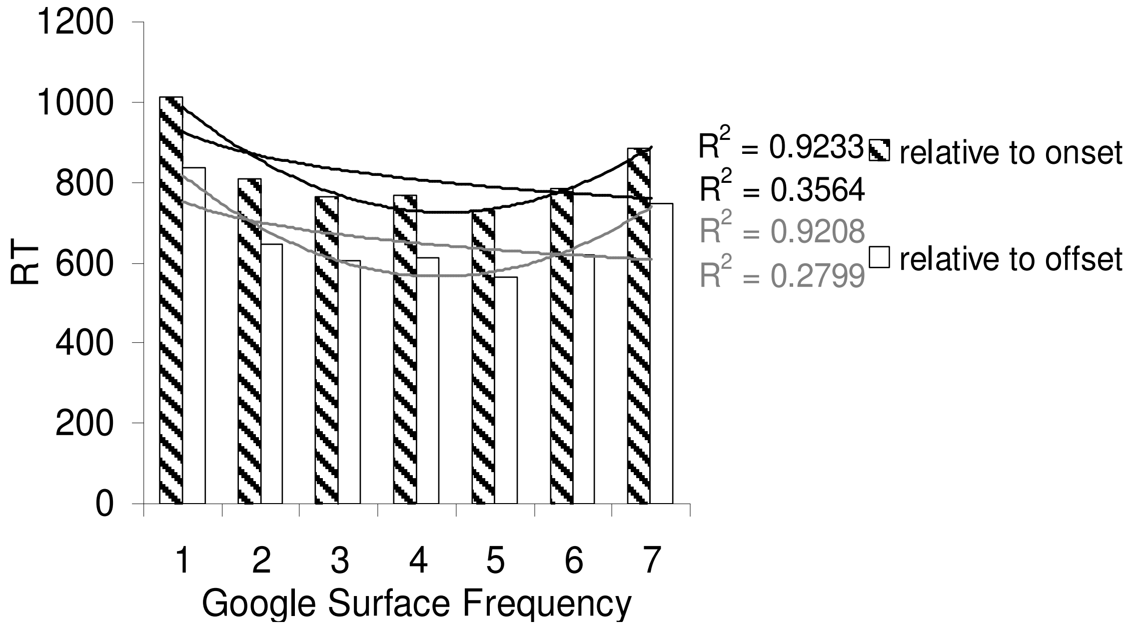
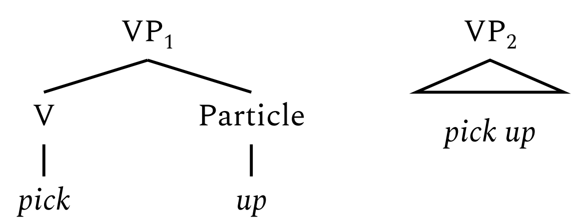


Figure 1.4.2.: A visualization of a holistically stored phrase with internal structure (left) and without internal structure (right).



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 (which is the only plausible explanation for how one starts with holistic chunks without resulting in holistic storage later down the line). Further, many high frequency and high predictability phrases have semantically vague relationships (e.g., *trick or treat*). These phrases may be difficult to break down into their component words due to the lack of semantic transparency. This may lead to the holistic storage of these phrases. However, it is possible that their representations for these items are updated to reflect knowledge of the meaning of the individual words upon learning the individual words. 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). Specifically, the authors demonstrated that given a novel semantic context, learners are more likely to use a frequent suffix to describe the novel context than an infrequent suffix. They argued that this is because the frequent suffix is more accessible. It is possible that semantic extension may also lead to a loss of internal structure in the multi-morphemic word (or phrase): as the 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 upon hearing part of the phrase, listeners may access the representation for the holistically stored phrase. For example, hearing *Habeas* may be enough to access the holistic 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 the interpretation at the first noun is implausible; *spread out the monkey* is not plausible. Sentences 2a and 2b are analogous but with a high frequency compound noun (where frequency is the number of times the phrase occurred, not the number of times the individual words did).

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 in 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 effect on recognition.

Following these results, it is possible that an incremental approach combined with competition (such as one proposed by McClelland et al., 1984) may be another possible 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 whether current learning theories necessarily predict holistic storage. In order to help examine this, we turn to large language models, which have developed rapidly over the last several years.

1.6.1. Transformer Model Architecture

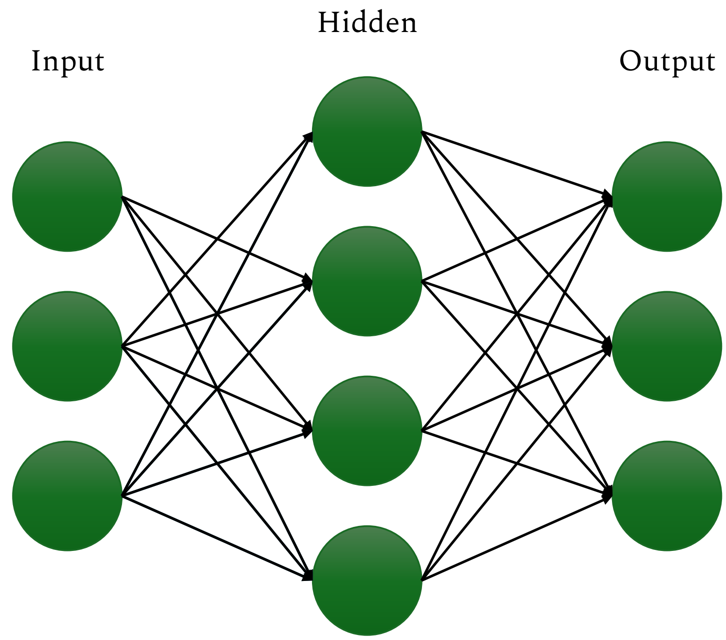
When I started this program in 2020, the idea of a single language model that could produce fluent text that was discernible from text written by humans seemed like a far-off dream. However, with rapid advancements in transformer models, the language models of today seem to have accomplished that goal. 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 trade off between stored and general 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). 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 implements a self-attention mechanism. The self-attention mechanism 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

[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 actually predicts (i.e., assigns the greatest probability to) *k* as the upcoming token in the example context.

Figure 1.6.1.: A visualization of a feed-forward neural network.



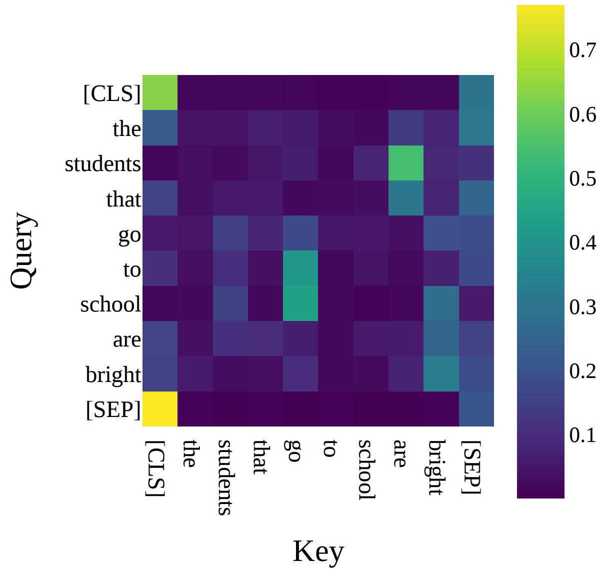
predecessors. For example, previous models such as Long-Short Term Memory (LSTM) models or Recurrent Neural Network (RNN) models struggled with long-term dependencies (Al-Selwi et al., 2023; Bengio et al., 1993). The self-attention mechanism was proposed as a solution to this limitation. The self-attention 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 are very 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*). This example is visualized in Figure 1.6.2 using BERT.

The full transformer model architecture is presented in Figure 1.6.3, 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 1.6.2.: A visualization of key-query attention values for BERT at layer 3, attention head 8. Notice that the strength between 'students' and 'are' is high. Brighter colors denote larger attention weights, darker colors denote smaller attention weights.

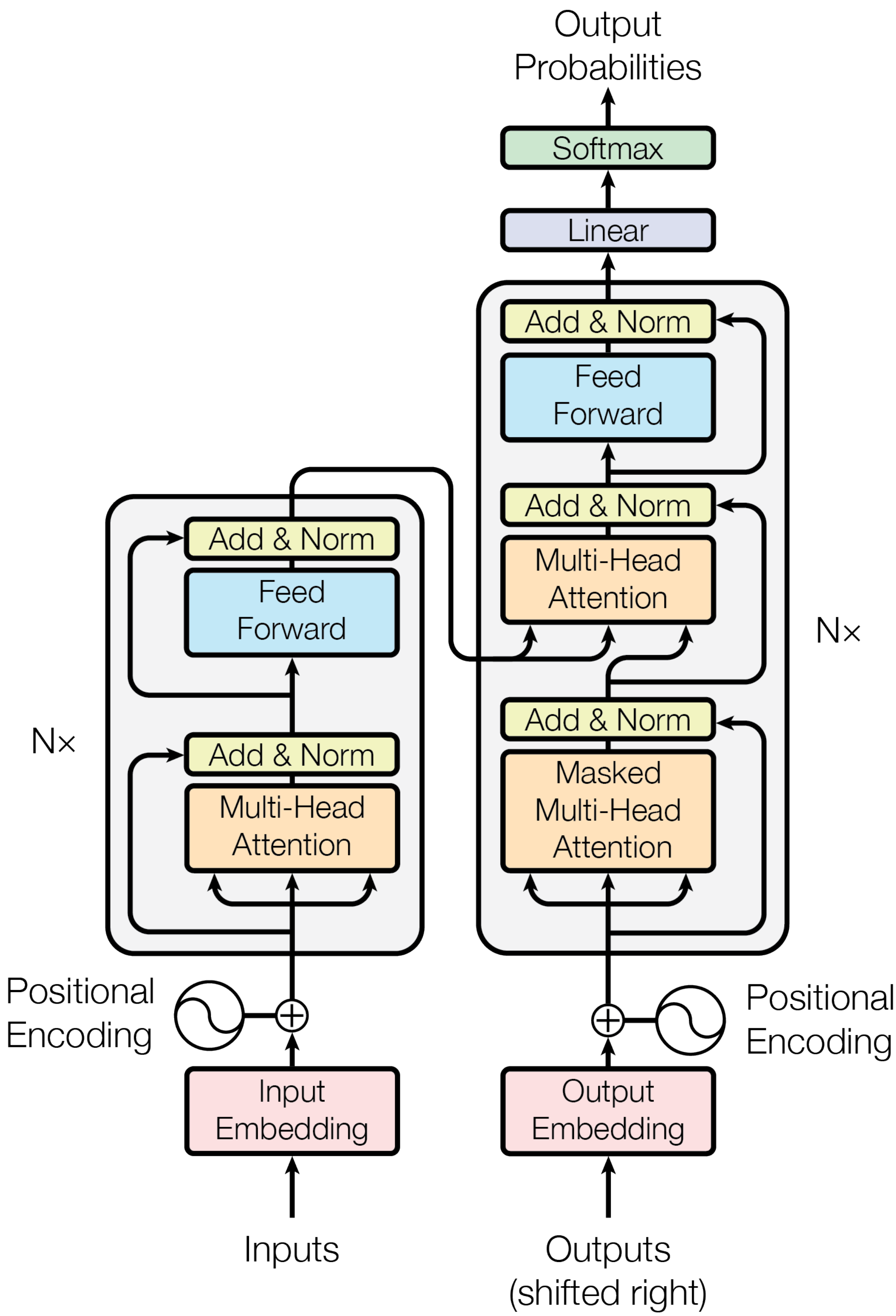


they do this in a human-like manner. Specifically, how much are these models simply memorizing as opposed to learning something more abstract about the language?

For example, there have been doubts about whether they're capable of learning anything abstract, such as meaning (e.g., Bender et al., 2021). Bender & Koller (2020) argued that language models cannot learn meaning because they are trained on only the form of the language. They operationalized meaning as the relationship between form and communicative intent. However, as Piantadosi & Hill (2022) stated in his rebuttal, this view of meaning ignores a well-known aspect of human language: meaning is not as simple as an association between a form and a referent. For example, the meaning *justice* has no real-world referent. Instead, people learn these abstract words through their relationship with other words. This type of learning is something that large language models do well (Piantadosi, 2023).

Additionally, while there have been many arguments that large language models don't learn in a human-like manner (Bender et al., 2021; Bender & Koller, 2020), it is rather unclear what is meant by this statement. On one hand it is the case that large language models are trained on trillions of tokens

Figure 1.6.3.: The transformer model architecture, reproduced from Vaswani et al. (2017).



(e.g., Groeneveld et al., 2024), which is magnitudes larger than humans who have heard an average of 350 million words by the time they enter college (Levy et al., 2012). This is further complicated 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. On the other hand, it is hard to compare the input that a language model to the input that humans receive. While it is true that the number of tokens that large language models receive is magnitudes larger than humans, humans receive a lot of contextual information when they hear a word in the form of sensory information from their environment. It could be the case that this makes learning in large language models impossible, or it could be the case that it just requires more data for them to learn. Given the performance of large language models, the latter case seems to be plausible.

Further, the argument that the learning mechanisms in humans and large language models are completely different also falls apart quite quickly. Large language models, as was demonstrated from the previous section, learn from prediction error. They update their representations to maximize prediction of the upcoming token (Vaswani et al., 2017). Learning theories (Kapatsinski, 2018, 2023; e.g., RESCORLA, 1968) have made the same argument for quite some time, demonstrating that humans and animals are sensitive to upcoming events. Further, the language processing literature has confirmed this prediction by demonstrating that humans are actively predicting upcoming linguistic information (Bansal et al., 2018; Clark, 2013; Ferreira & Chantavarin, 2018; Kuperberg & Jaeger, 2016; Olejarczuk et al., 2018; Ramscar et al., 2013). This isn't to say that language models are learning identically to humans, but rather that there is enough overlap between large language models and humans to warrant a close examination of them. In fact, learning the differences between humans and large language models may be a fruitful endeavor, leading to improvements in both language models but also theories of language in humans.

Thus in this section we take a closer examination into what evidence there is that language models are learning something abstract as opposed to simply memorizing.

First, 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. Specifically, he tested BERT on English, French, Dutch, and Spanish on a number-agreement task. He evaluated whether the model was able to predict the correct plural form of novel-words in a no-prime condition and a prime-condition, where a previous sentence reveals the correct form of the verb (as well as the appropriate gender for the languages with a gender distinction). Humans are able to reliably use the prime to form the correct plural even for non-words. Interestingly, they found that BERT was not able to reliably use the prime to improve performance on the non-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. They examined the ability of BERT to learn the correct tense in French and Chinese and found that while BERT is able to do well in French, it does much worse in Chinese. They argued that this is because while in French the correct tense information is expressed in the verb morphology (and can therefore be predicted by surface statistics of the language), in Chinese tense is driven by a variety of different cues, including abstract, lexical, syntactic, and pragmatic information. The poor performance in Chinese thus, according to the authors, suggests that BERT relies on memorization of surface-level statistics.

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 used a probing task to determine whether the model learned anything general about the language. The probing task attempted to predict the number of the verb and the object-past agreements from the representations in the model. Their probe achieved a high accuracy, suggesting that the model encoded an abstract representation for these linguistic features.

The evidence for abstractions is not limited to BERT, either. 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.

In addition to large language models, 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. Finally, Chapter 6 examines whether storage accounts can explain frequency-dependent preference extremity.

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, 2015, 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 (Bybee, 2003, pp. 14–15).

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.

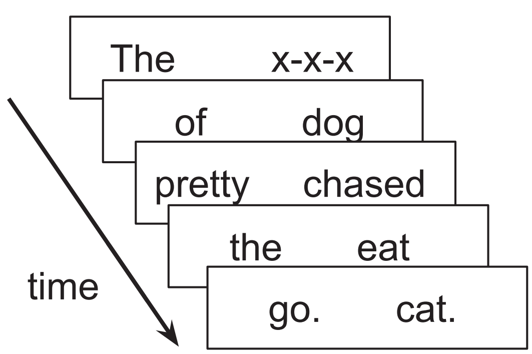
Stimuli

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. (2018). 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 for a visualization of the maze task, reproduced from Boyce et al. (2020).

Figure 2.2.1.: 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 2.2.1 and in Figure 2.2.2, and the results of the N2 region are presented in Table 2.2.2 and in Figure 2.2.3.

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

Table 2.2.1.: Model results examining the effect of plausibility and frequency for the N1 region.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	6.82	0.02	6.78	6.87	100.00
Plausibility	0.06	0.01	0.04	0.08	83.86
Familiarity	0.01	0.01	-0.01	0.04	100.00
Plausibility:Familiarity	0.00	0.01	-0.02	0.02	38.76

Table 2.2.2.: Model results examining the effect of plausibility and frequency for the N2 region.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	6.87	0.03	6.82	6.92	100.00
Plausibility	-0.07	0.01	-0.08	-0.05	0.06
Familiarity	-0.07	0.02	-0.11	-0.03	0.00
Plausibility:Familiarity	0.00	0.01	-0.01	0.02	63.36

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.

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

Figure 2.2.2.: Plot of log reaction time at the N1 region as a function of plausibility and familiarity.

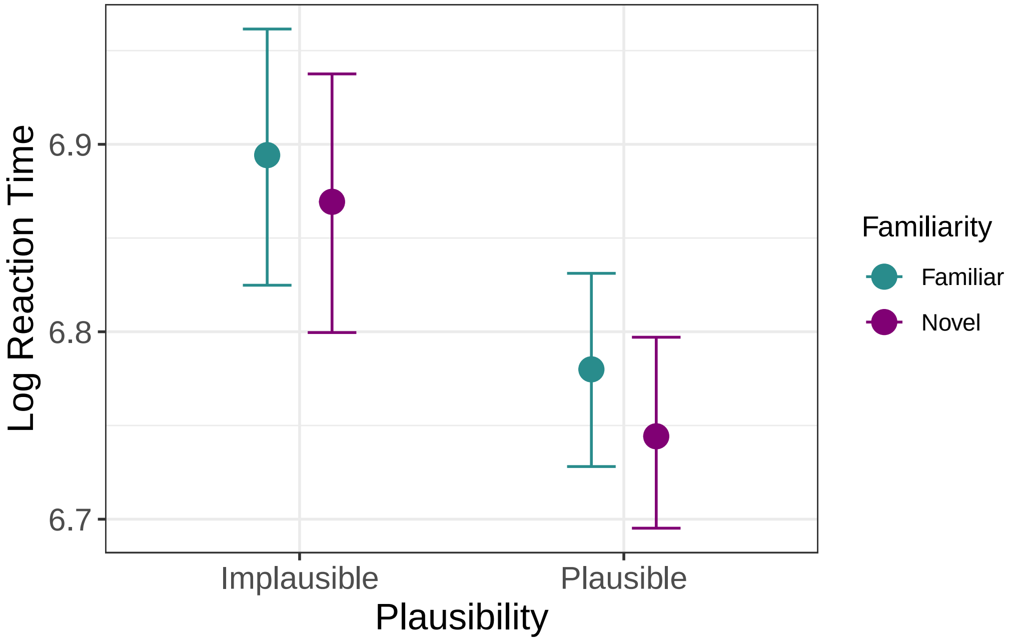
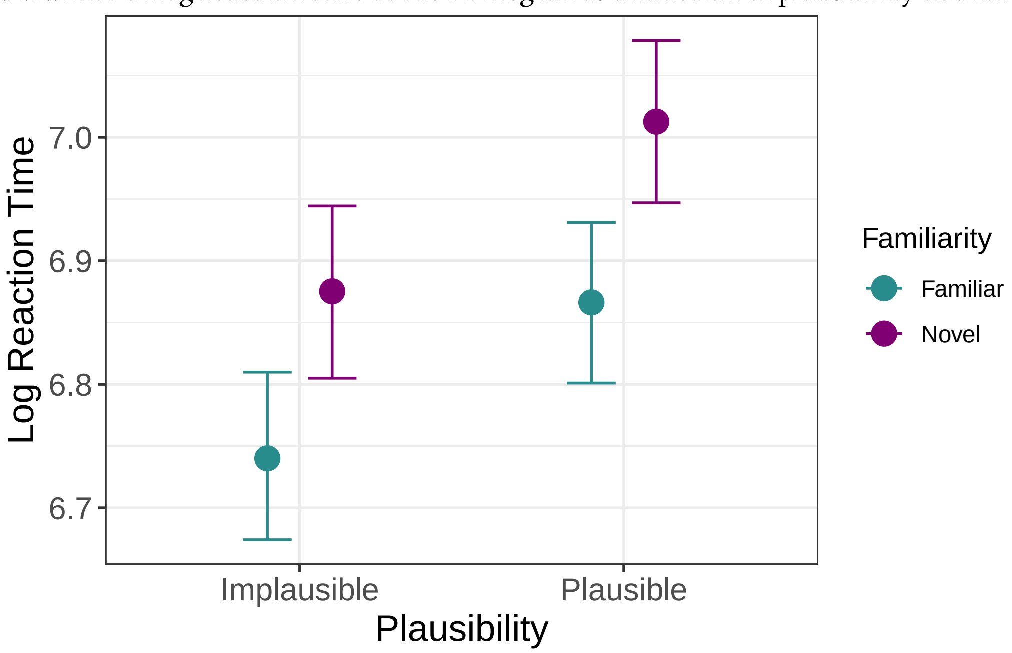


Figure 2.2.3.: Plot of log reaction time at the N2 region as a function of plausibility and familiarity.



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**

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).

[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.

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., 2018). 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 2.3.1 and Table 2.3.2, and visualized in Figure 2.3.1 and Figure 2.3.2. The results of the N2 region are presented in Table 2.3.3 and Table 2.3.4, and visualized in Figure 2.3.3 and Figure 2.3.4.

Table 2.3.1.: 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	6.88	0.03	6.82	6.94	100.00
Plausibility	0.07	0.01	0.04	0.10	100.00
Predictability	0.03	0.02	-0.01	0.08	92.30
Plausibility:Predictability	0.00	0.01	-0.03	0.03	46.72

Table 2.3.2.: 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	6.88	0.03	6.82	6.93	100.00
Plausibility	0.07	0.01	0.04	0.10	100.00
LogOdds	0.00	0.01	-0.01	0.02	76.50
Plausibility:LogOdds	0.00	0.00	-0.01	0.01	61.64

With regards to the N1 region, Table 2.3.1 presents the results of the analysis we ran with predictability as a binary predictor (high or low), while Table 2.3.2 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 2.3.3 presents the results of the analysis we ran with predictability as a binary predictor (high or low), while Table 2.3.4 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 2.3.1 and Figure 2.3.3 provide visualizations of the analyses run with predictability as a binary variable while Figure 2.3.2 and Figure 2.3.4 present analyses with predictability as a continuous variable.

Table 2.3.3.: 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	6.79	0.03	6.74	6.85	100.00
Plausibility	-0.07	0.01	-0.10	-0.05	0.00
Predictability	-0.08	0.03	-0.13	-0.03	0.12
Plausibility:Predictability	0.00	0.01	-0.02	0.03	67.46

Table 2.3.4.: 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	6.80	0.03	6.75	6.85	100.00
Plausibility	-0.07	0.01	-0.10	-0.05	0.00
LogOdds	-0.03	0.01	-0.05	-0.02	0.00
Plausibility:LogOdds	0.00	0.00	-0.01	0.01	60.44

Figure 2.3.1.: Plot of the N1 region with predictability as a binary variable (high or low).

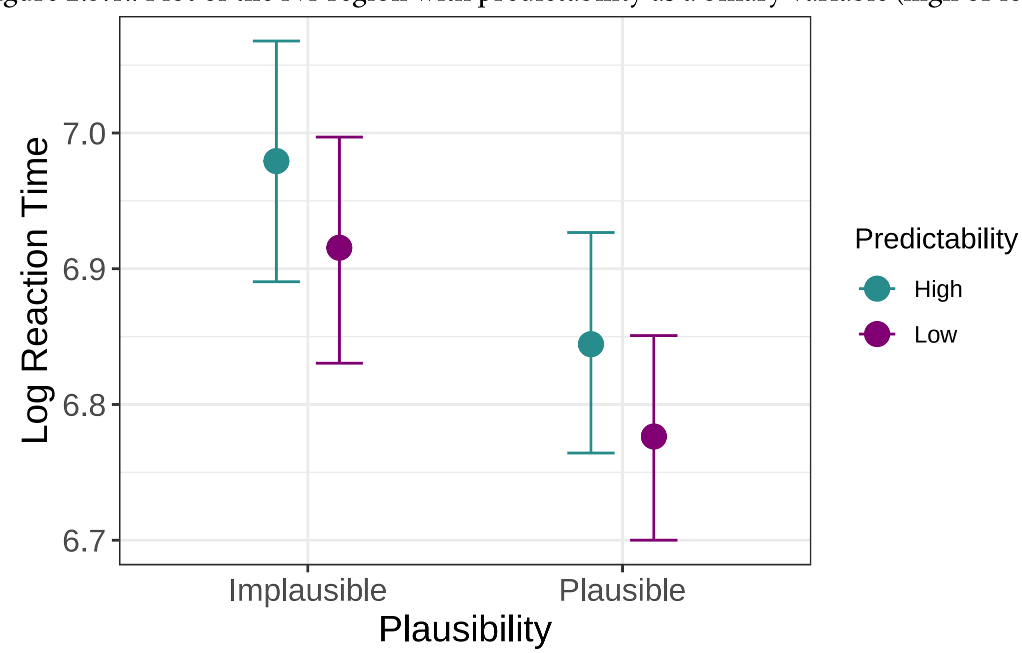


Figure 2.3.2.: Plot of the N1 region with predictability as a continuous variable.

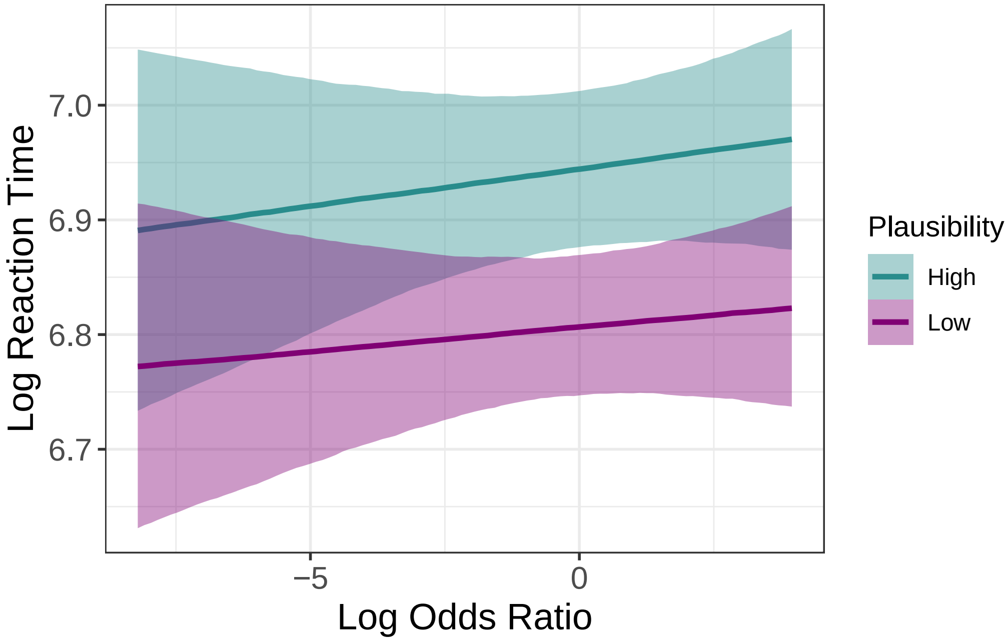


Figure 2.3.3.: Plot of the N2 region with predictability as a binary variable (high or low).

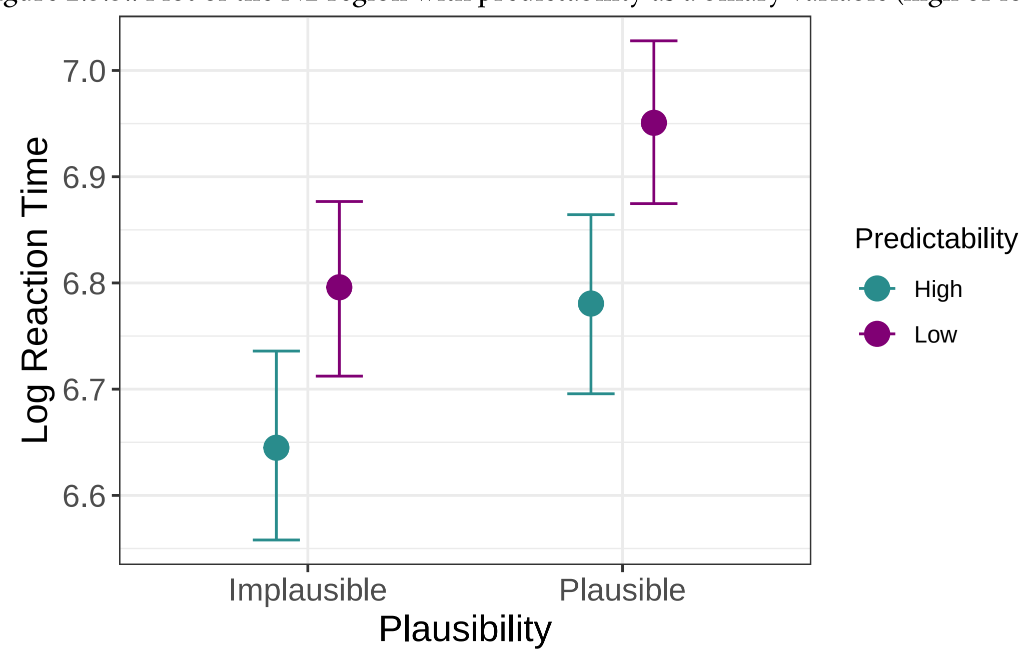
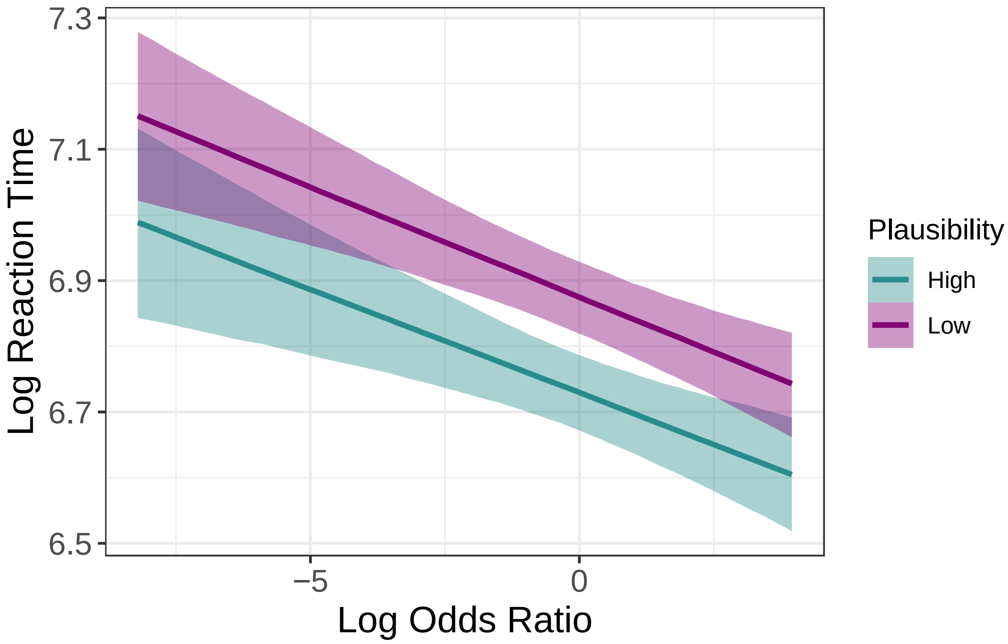


Figure 2.3.4.: Plot of the N2 region with predictability as a continuous variable.



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.

Table 2.4.1.: Model results for each eye-tracking measure at the N1 region.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
First Fixation Duration					
Intercept	239.23	5.54	228.46	250.05	100.00
Plausibility	4.19	2.16	-0.02	8.38	97.40
Familiarity	-0.87	2.40	-5.59	3.78	35.52
Plausibility:Familiarity	0.03	2.40	-4.84	4.78	49.68
Gaze/First-Pass Duration					
Intercept	273.96	8.56	257.09	290.93	100.00
Plausibility	0.04	0.20	-0.33	0.42	57.83
Familiarity	0.00	0.20	-0.40	0.38	49.73
Plausibility:Familiarity	0.01	0.20	-0.38	0.40	51.92
Go-Past Time					
Intercept	363.00	18.58	325.95	399.74	100.00
Plausibility	16.90	6.91	3.61	30.54	99.37
Familiarity	-3.75	7.89	-19.44	11.85	31.34
Plausibility:Familiarity	14.84	7.42	0.45	29.41	97.77
First-Pass Regression					
Intercept	-1.99	0.18	-2.35	-1.64	0.00
Plausibility	0.16	0.09	-0.01	0.33	96.38
Familiarity	0.02	0.09	-0.15	0.19	57.80
Plausibility:Familiarity	0.06	0.09	-0.12	0.23	74.50

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

N1 Region

Our results at the N1 region are demonstrated in Table 2.4.1 and visualized in Figure 2.4.1.

At the N1 region, we find main-effects of plausibility for first fixation duration, go-past time, and first-pass regression, but not for gaze times. Additionally, we find no effects of familiarity. Finally, for go-past times we find an interaction between plausibility and predictability such that the slowdown of the implausible context was greater for familiar items than novel items.

Figure 2.4.1.: Visualization of the effects of plausibility and familiarity on each eye-tracking measure at the N1 region.

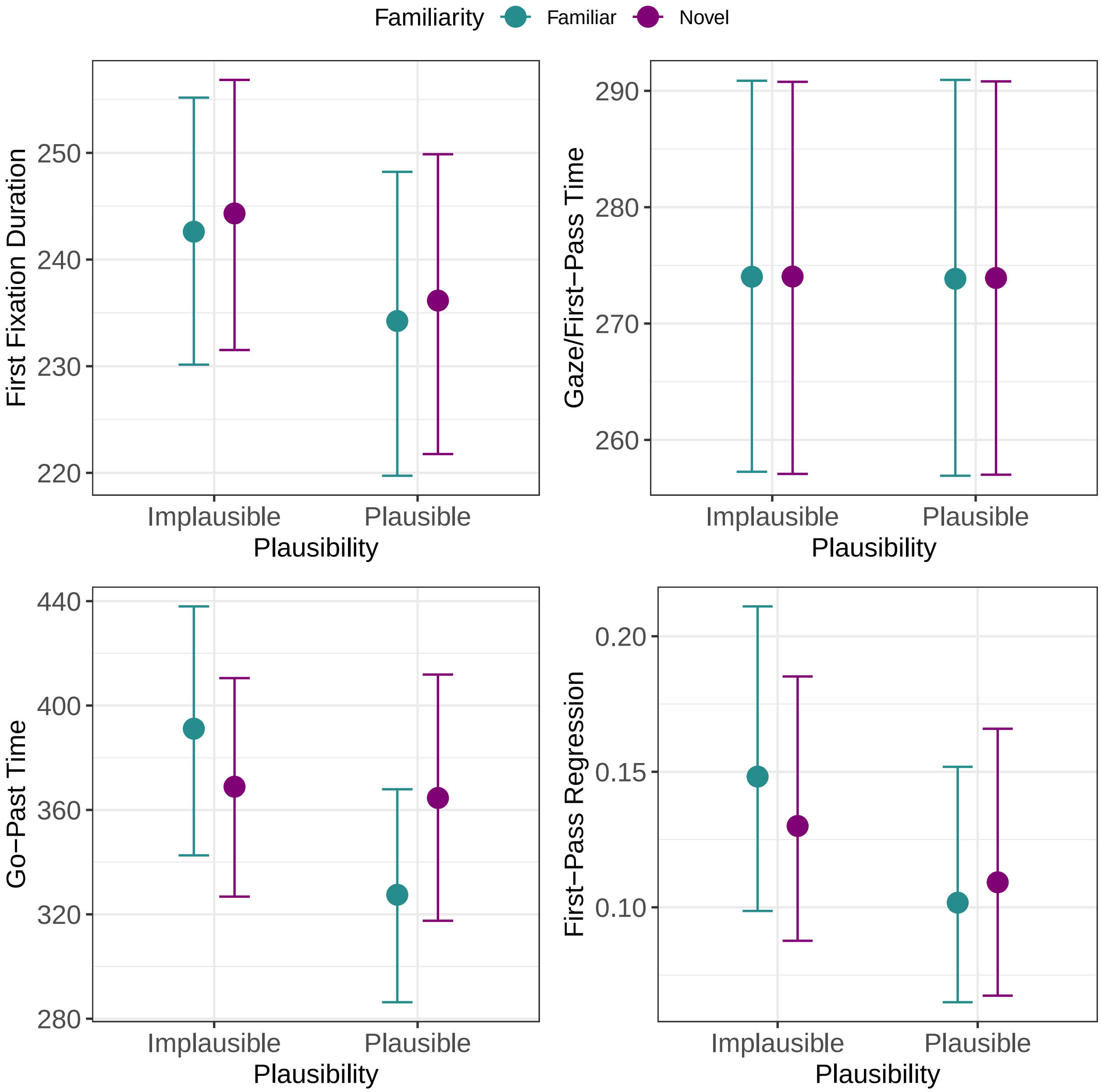


Table 2.4.2.: Results of models for each eye-tracking measure at the N2 region.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
First Fixation Duration					
Intercept	249.62	6.47	236.38	262.35	100.00
Plausibility	-2.09	2.65	-7.40	3.12	21.58
Familiarity	-6.97	4.27	-15.43	1.44	5.42
Plausibility:Familiarity	-1.05	2.37	-5.67	3.62	32.48
Gaze/First-Pass Duration					
Intercept	275.48	9.04	257.70	293.10	100.00
Plausibility	0.05	3.76	-7.50	7.44	50.55
Familiarity	-12.96	6.37	-24.98	-0.20	2.33
Plausibility:Familiarity	-3.88	3.78	-11.31	3.42	15.24
Go-Past Time					
Intercept	351.82	21.88	308.49	394.20	100.00
Plausibility	6.95	8.40	-9.27	23.77	79.87
Familiarity	-17.43	13.36	-44.88	7.49	8.97
Plausibility:Familiarity	-7.96	9.35	-26.62	10.25	19.43
First-Pass Regression					
Intercept	-2.31	0.21	-2.75	-1.91	0.00
Plausibility	0.07	0.09	-0.10	0.26	78.45
Familiarity	-0.14	0.12	-0.37	0.10	11.58
Plausibility:Familiarity	0.00	0.09	-0.18	0.17	48.58

N2 Region

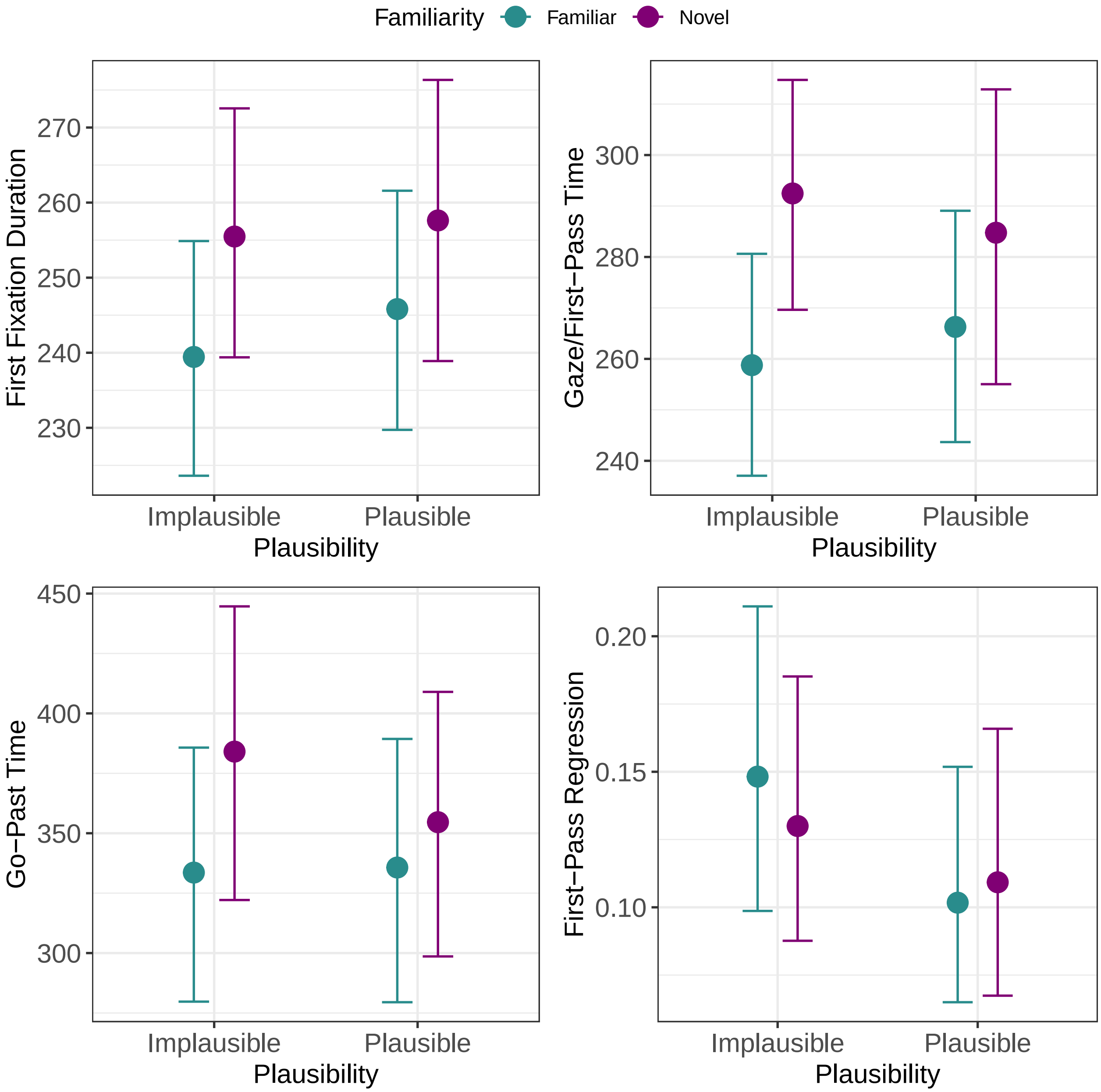
Our results at the N2 region are demonstrated in Table 2.4.2 and visualized in Figure 2.4.2.

At the N2 region, we find main-effects of familiarity for first fixation duration and gaze times, and marginal effects for go-past times and first-pass regression. We find no main-effect of plausibility, which is expected because the implausible condition is plausible at the N2 region. We also find no interaction effect between plausibility and familiarity.

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.

Figure 2.4.2.: Visualization of the effects of plausibility and familiarity on each eye-tracking measure at the N2 region.



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

N1 Region

Our results at the N1 region are demonstrated in Table 2.5.1 and visualized in Figure 2.5.1.

At the N1 region, we find main-effects of plausibility for only first-pass regression. We find no effect of plausibility for first fixation duration, gaze time, and go-past time. Additionally, we find no effects of predictability. Finally, we find an interaction effect between plausibility and predictability for first fixation duration and first-pass regression. These interaction effects are in the opposite direction such that the slowdown caused by the implausible condition was greater for high-predictability items relative to low-predictability items in first-pass regression, but smaller for high-predictability items relative to low-predictability items in first fixation duration.

N2 Region

Our results at the N2 region are demonstrated in Table 2.5.2 and visualized in Figure 2.5.2.

Figure 2.5.1.: Visualization of the effects of plausibility and predictability on each eye-tracking measure at the N1 region.

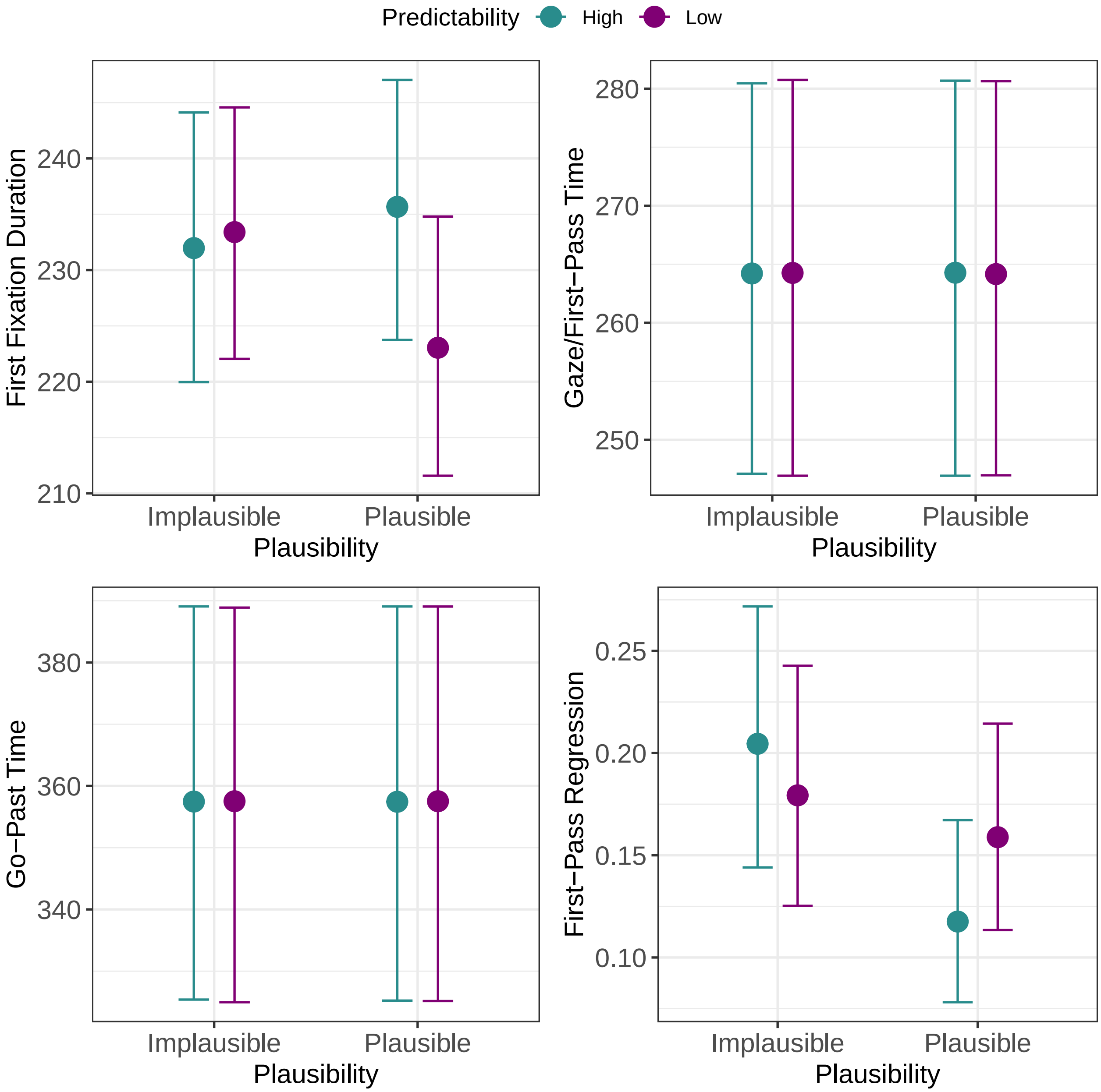


Table 2.5.1.: Model results for each eye-tracking measure at the N1 region.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
First Fixation Duration					
Intercept	231.02	4.43	222.36	239.50	100.00
Plausibility	1.66	2.02	-2.29	5.70	78.95
Predictability	2.78	2.87	-2.80	8.53	83.75
Plausibility:Predictability	-3.47	2.01	-7.42	0.52	4.15
Gaze/First-Pass Duration					
Intercept	264.14	8.42	246.90	280.60	100.00
Plausibility	0.00	0.20	-0.38	0.41	49.08
Predictability	0.00	0.20	-0.39	0.39	51.48
Plausibility:Predictability	-0.01	0.20	-0.41	0.38	48.10
Go-Past Time					
Intercept	357.21	16.45	325.20	388.96	100.00
Plausibility	0.02	0.20	-0.37	0.40	54.05
Predictability	0.00	0.20	-0.39	0.39	50.10
Plausibility:Predictability	0.00	0.20	-0.39	0.39	49.77
First-Pass Regression					
Intercept	-1.65	0.15	-1.95	-1.36	0.00
Plausibility	0.20	0.08	0.04	0.36	99.38
Predictability	-0.05	0.09	-0.21	0.12	27.75
Plausibility:Predictability	0.13	0.07	-0.02	0.28	96.03

At the N2 region, we find a main-effect of predictability for only the first fixation duration measure. We find no effect of plausibility, as expected because the N2 region eliminates the implausibility effect. We also find no interaction between plausibility and predictability in any of the measures.

Filler Items

In addition to our experimental stimuli, we also included sentences that varied with respect to the frequency of the word. For example, in the below sentences, *satchel* is low-frequency and *account* is high-frequency. Thus, in order to confirm that the results in Experiment 4 were not due to measurement error, we examined the effect of frequency in our filler items on each of the eye-tracking measures.

1. **Low-frequency**

    1 Take your money out of the satchel and pay off the debt.

2. **High-frequency**

Figure 2.5.2.: Visualization of the effects of plausibility and predictability on each eye-tracking measure at the N2 region.

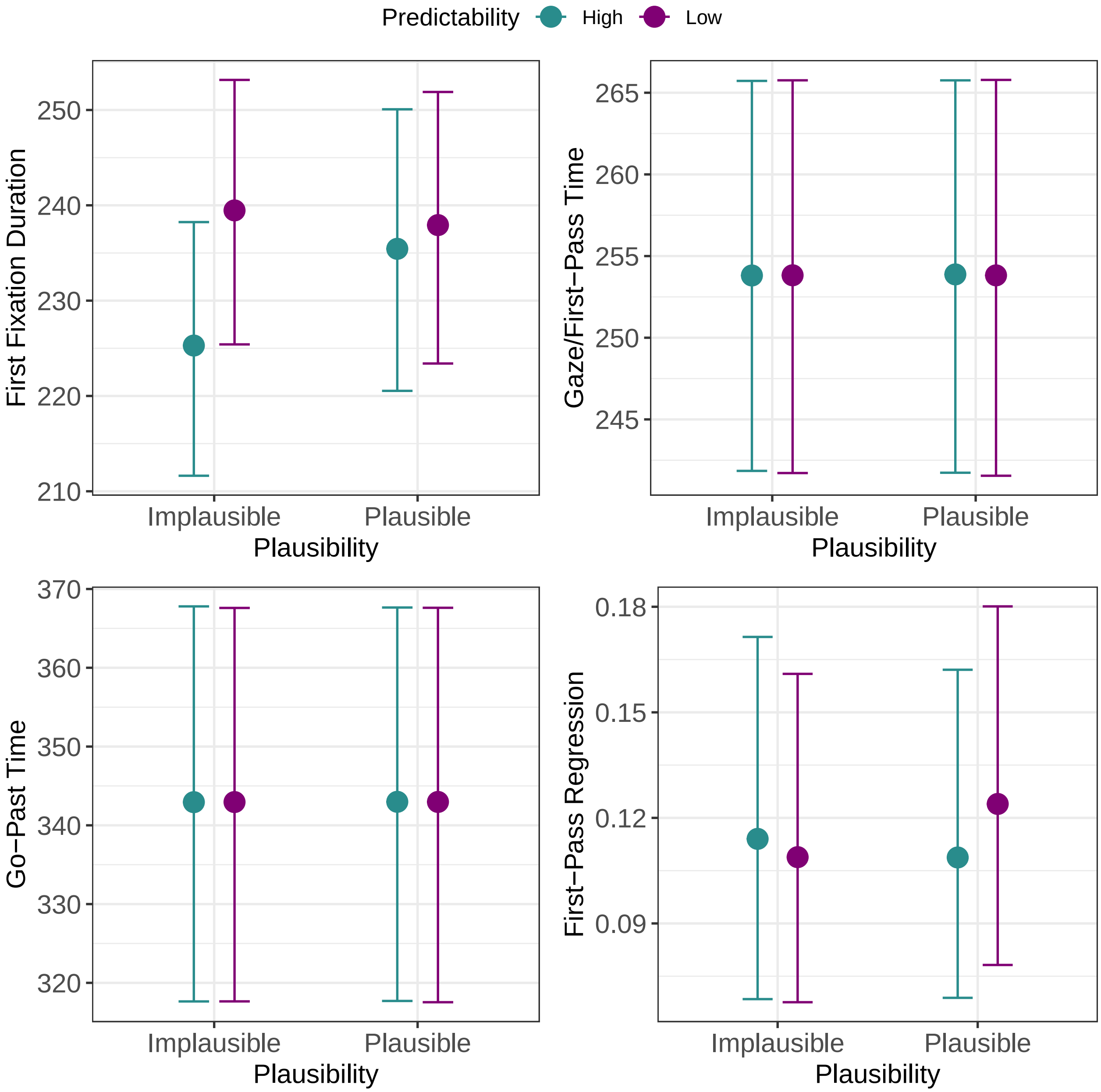


Table 2.5.2.: Model results for each eye-tracking measure at the N2 region.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
First Fixation Duration					
Intercept	234.45	4.86	224.48	243.77	100.00
Plausibility	-2.15	2.97	-8.13	3.68	23.23
Predictability	-4.14	2.99	-9.89	1.89	8.58
Plausibility:Predictability	-2.93	3.01	-8.92	2.97	16.08
Gaze/First-Pass Duration					
Intercept	253.76	6.07	241.73	265.77	100.00
Plausibility	0.00	0.10	-0.20	0.19	48.83
Predictability	0.00	0.10	-0.21	0.19	48.62
Plausibility:Predictability	0.00	0.10	-0.20	0.20	51.62
Go-Past Time					
Intercept	342.74	12.87	317.68	367.65	100.00
Plausibility	0.00	0.10	-0.20	0.20	47.60
Predictability	0.00	0.10	-0.20	0.20	50.00
Plausibility:Predictability	0.00	0.10	-0.19	0.20	49.02
First-Pass Regression					
Intercept	-2.06	0.18	-2.43	-1.73	0.00
Plausibility	-0.03	0.11	-0.24	0.18	40.62
Predictability	-0.02	0.10	-0.22	0.18	41.10
Plausibility:Predictability	0.05	0.10	-0.14	0.24	69.97

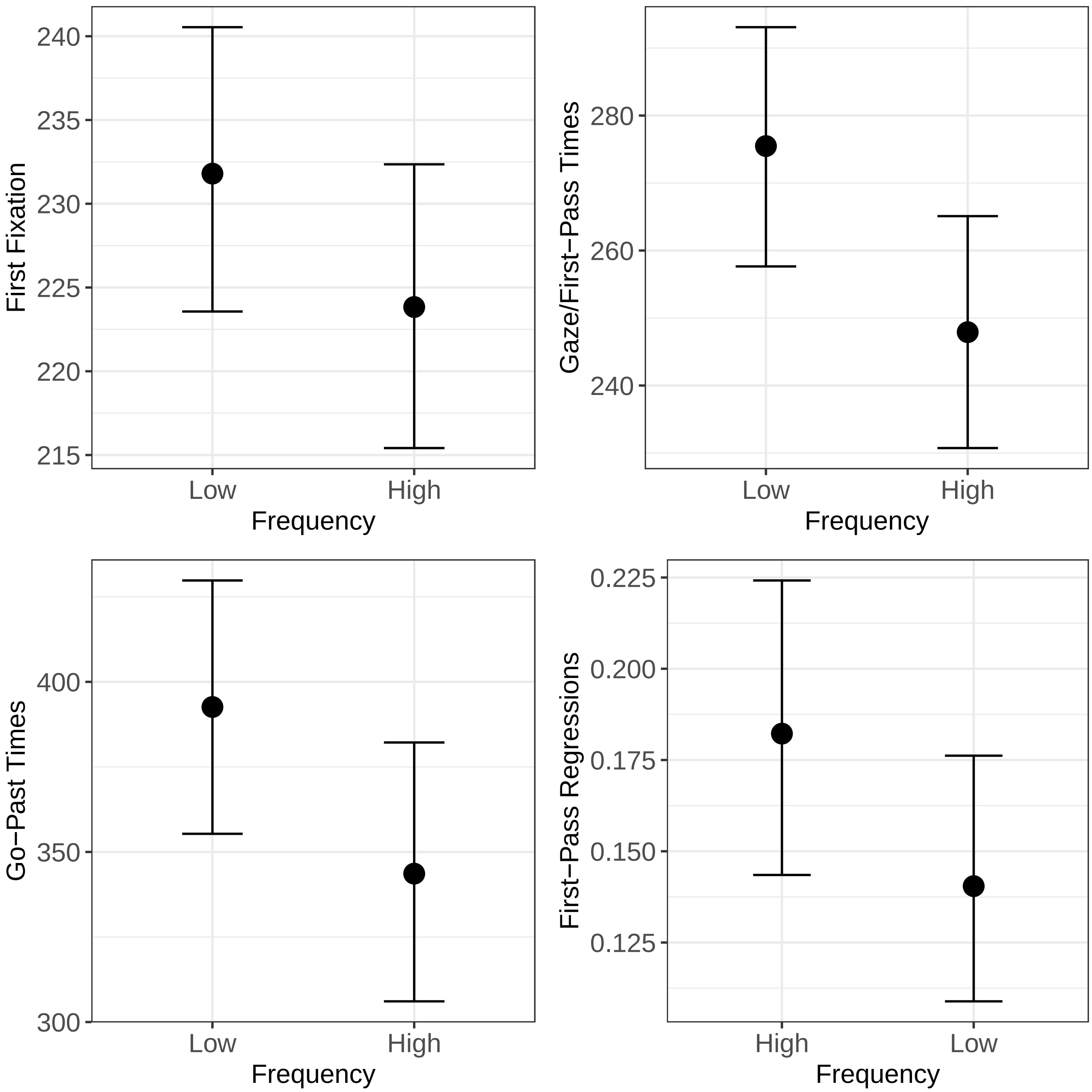
2 Take your money out of the account and pay off the debt.

Our results are presented in Table 2.5.3 and visualized in Figure 2.5.3. We find an effect of frequency in each of the four eye-tracking measures we looked at.

Table 2.5.3.: Model results for filler items for each eye-tracking measure

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
First Fixation Duration					
Intercept	227.843	3.942	220.186	235.404	100.00
Frequency	4.011	1.848	0.414	7.642	98.50
Gaze/First-Pass Duration					
Intercept	261.710	7.805	246.338	276.992	100.00
Frequency	13.825	4.371	5.179	22.322	99.88
Go-Past Time					
Intercept	368.199	17.340	334.329	402.836	100.00
Frequency	24.391	7.863	9.001	39.509	99.80
First-Pass Regression					
Intercept	-1.659	0.129	-1.921	-1.417	0.00
Frequency	0.156	0.058	0.044	0.271	99.60

Figure 2.5.3.: Visualization of the effects of frequency on each eye-tracking measure for filler items.



2.5.3. Discussion

In Experiment 4, we find an effect of plausibility only in first-pass regressions. Interestingly, we do find an effect of plausibility in first-fixation times for low-predictability items, but not for high-predictability items. While this follows our theoretical prediction that high-predictability items may be able to overcome the local implausibility, the results are difficult to reconcile with the results we see for first-pass regressions. For first-pass regressions, we observe the opposite pattern: 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 study 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 found an effect of plausibility in first fixation times, go-past times, and first-pass regressions. We also found 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 frequency or predictability of the second noun in the compound (given the first noun) has very little facilitatory effect on the processing of the first noun in implausible contexts relative to plausible contexts. That is, the increase in reaction time in the implausible condition for the N1 region was not mediated by the frequency or 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 because the second noun always eliminated the local implausibility.

Experiments 3 and 4 provide mixed-evidence. On one hand, there seems to be a general effect of implausibility regardless of frequency or predictability, however in some reading measures such as first-fixation times in Experiment 4, predictability seemed to alleviate the slowdown in the implausible condition. On the other hand, for other measures, the slowdown generated by the implausible contexts was actually exacerbated for high-frequency or high-predictability items (e.g., first-pass regression times in both Experiment 3 and Experiment 4).

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.

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, 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 may 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 during sentence processing, readers do not seem to access the holistic representation at the first noun (because then they would be able to overcome the implausibility). Instead, it is possible that if compound nouns are stored holistically, perhaps readers don't access the holistic representation until they have heard sufficient evidence for that representation. That is, even though *peanut* is predictive of *butter*, there are still many other words that can occur after *peanut*. Thus, readers may not access the holistic representation until they hear enough of the compound noun to rule out other competing possibilities.

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 3.1.2). 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 3.1.1). 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 3.1.2. 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 3.1.1.: The U-shaped effect of the frequency of verb+*up* constructions on the speed with which *up* is detected, reproduced from Kapatsinski & Radicke (2009).

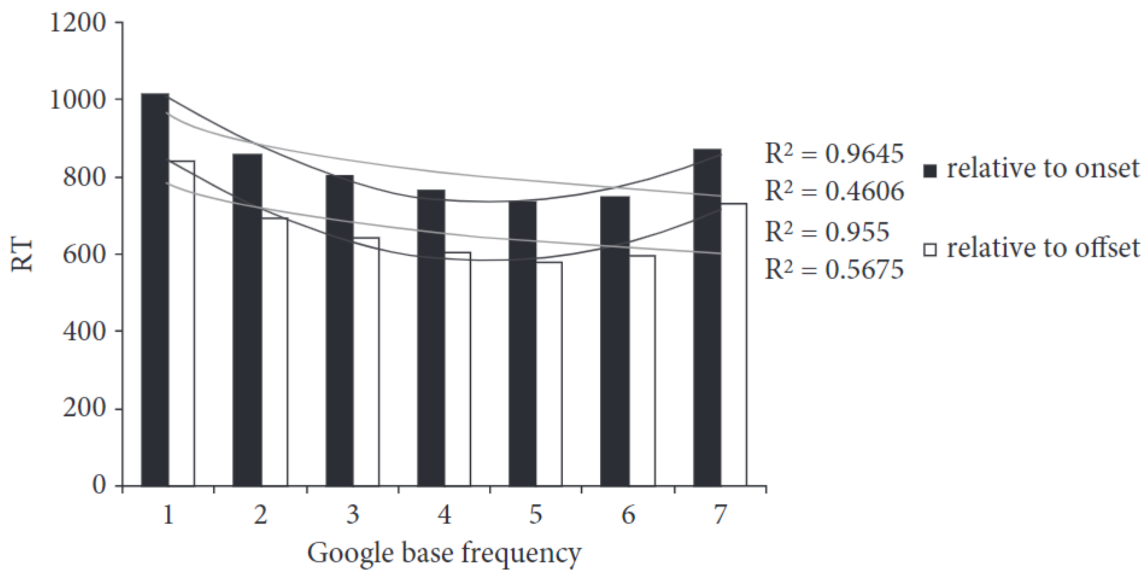
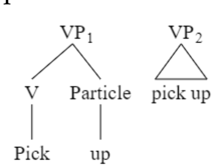


Figure 3.1.2.: A diagram of two ways the word *pick up* could be stored. The left tree demonstrates a stored representation of *pick up*, where the internal structure is still intact. The right tree demonstrates a holistically stored unit, where there is a loss of internal structure. Note that both of 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).

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

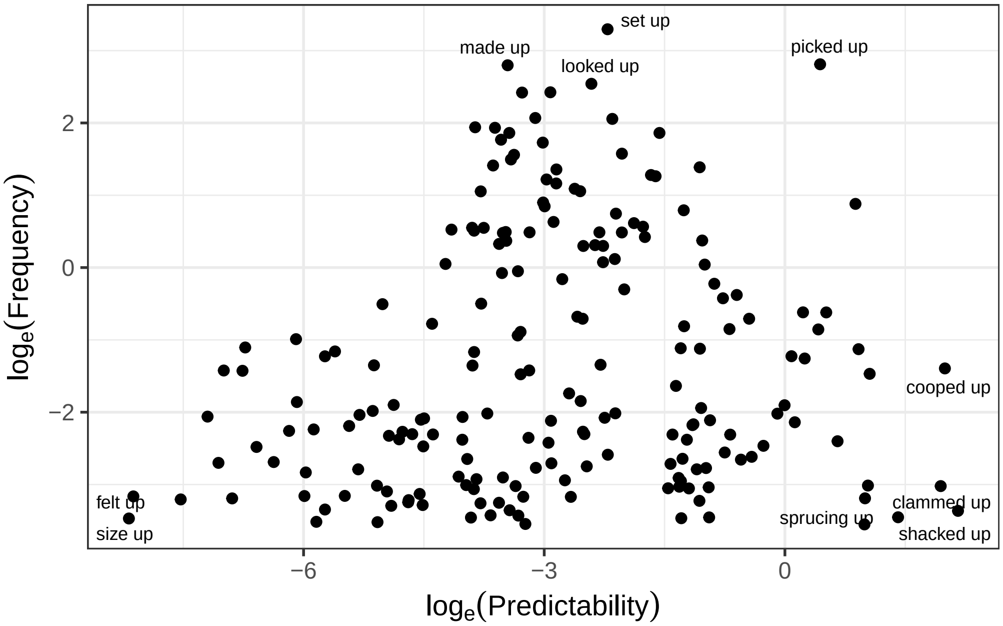
In non-mathematical terms, the above equation quantifies how likely *up* is to follow after the verb relative to every other word that could follow. For example, the odds ratio of *pick up* would be the number of times the entire verb phrase occurs – *pick up* – divided by the number of times the verb – *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 3.2.1.

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 3.2.1.: 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.

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

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

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, λ, 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 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,

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

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 3.3.1 and visualized in Figure 3.3.1. 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 obscuring the interaction effect. Thus, we ran an additional model examining whether the interaction effect was different for phrasal verbs versus non-phrasal verbs. The model equation is included below in Equation 3.3:

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

Our results for this model are reported in Table 3.3.2 and visualized in Figure 3.3.2. Over-

[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.

all our results replicate the results from the the model that didn't include phrasal verb as a predictor (Equation 3.2). Specifically, our results suggest that there is no interaction effect between frequency and predictability for phrasal verbs and non-phrasal verbs alike.

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 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 3.4:

$$\begin{aligned} log(RT) \sim\ & ti(Predictability) + ti(Frequency) + ti(Predictability, Frequency) \\ & + Duration + s(participant, bs = \text{`}re\text{'}) + s(Item, bs = \text{`}re\text{'}) + s(trial, bs = \text{`}re\text{'}) \\ & + s(Predictability, Frequency, Trial, Participant, bs = \text{`}re\text{'}) \end{aligned} \tag{3.4}$$

Our results are presented in Table 3.3.3 and visualized in Figure 3.3.3. The results demonstrated a significant main-effect of predictability ($p < 0.05$), but no significant effect of frequency ($p = 0.327$), and no significant interaction effect.[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

[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.

that there was no significant interaction between frequency and predictability, we left out the interaction term from the regression model. Specifically, we modeled log RT as a function of log-frequency, log-predictability, log-frequency^2, log-predictability^2, and duration. We also included maximal random effects structure (following Barr et al., 2013). The random-effects were modeled without correlations between them in order to allow the model to run faster. Equation 3.5 below presents the full model syntax:

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

The results of this model are presented in Table 3.3.4 and visualized in Figure 3.3.4. 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 3.3.5. 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 3.6 and Equation 3.7:

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

$$\begin{aligned} log(RT) &\sim log(Predictability) + Duration + log(Predictability)^2 \\ &+ (1 + log(Predictability) + log(Predictability)^2 + Duration||Participant) \\ &+ (1||Item) \end{aligned} \tag{3.7}$$

The results of our first model are presented in Table 3.3.5. 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. Specifically, we find a main-effect of of log(frequency)^2 ($\beta = 0.006$) comparable to the effect from our full quadratic model (Equation 3.5, $\beta = 0.005$).

The results of our second model are presented in Table 3.3.6. While the credible interval for log(predictability)^2 crosses zero, over 96% of the posterior samples were greater than zero, suggesting a meaningful effect. Specifically, we find a main-effect of log(predictability)^2 ($\beta = 0.003$) comparable to

Table 3.3.1.: 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 3.3.2.: 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

the effect from our full quadratic model model (Equation 3.5, $\beta = 0.003$). In other words, the results from both of our individual quadratic regression models (Equation 3.6 and Equation 3.7) replicate those found in Table 3.3.4.

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. Finally, we do not find strong evidence for an interaction between frequency and predictability regardless of whether the item was a phrasal verb or not, but it is possible that our study simply does not have the power to detect an interaction effect.

Table 3.3.3.: 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 3.3.4.: 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

Table 3.3.5.: 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 3.3.6.: 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

Figure 3.3.1.: 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.

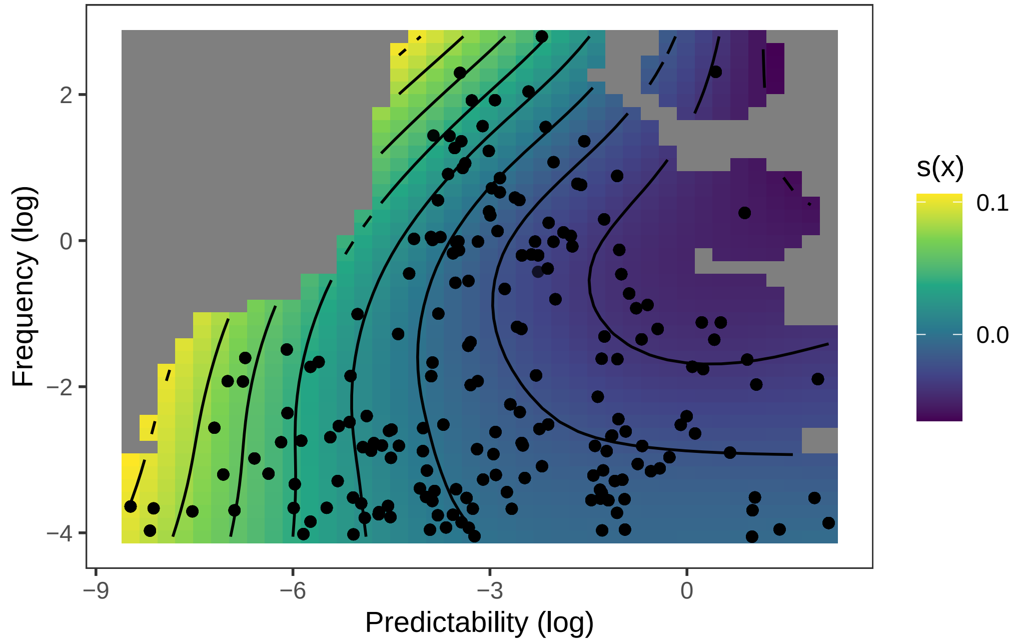


Figure 3.3.2.: Plot of the interaction effect between predictability and frequency of our GAM model containing the interaction between frequency and predictability for phrasal vs non-phrasal verbs. Brighter colors denote longer reaction times. The left graph is the predicted effect for phrasal verbs (e.g., pick up), the right graph is the predicted effect for non-phrasal verbs (e.g., walk up).

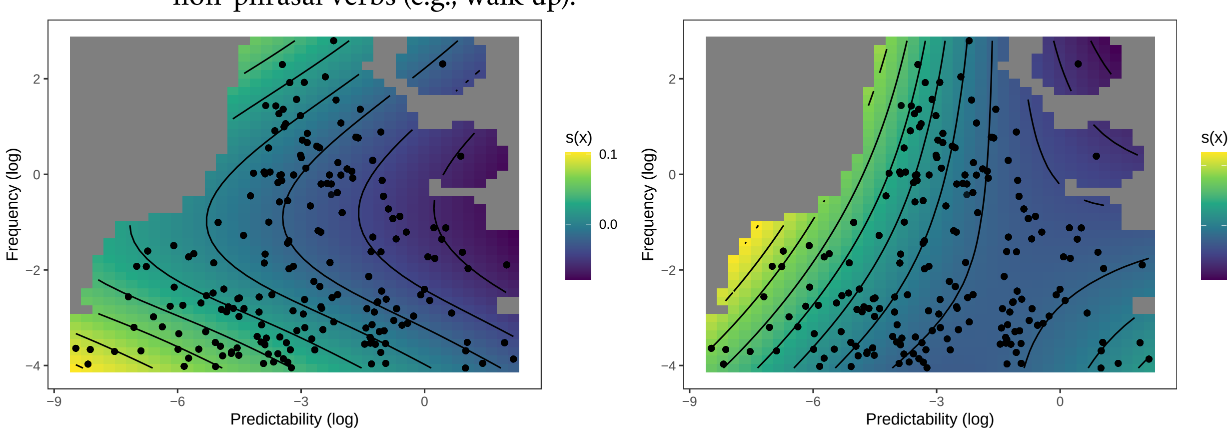


Figure 3.3.3.: Plot of our GAM model's predicted effect of log(predictability) on recognition time.

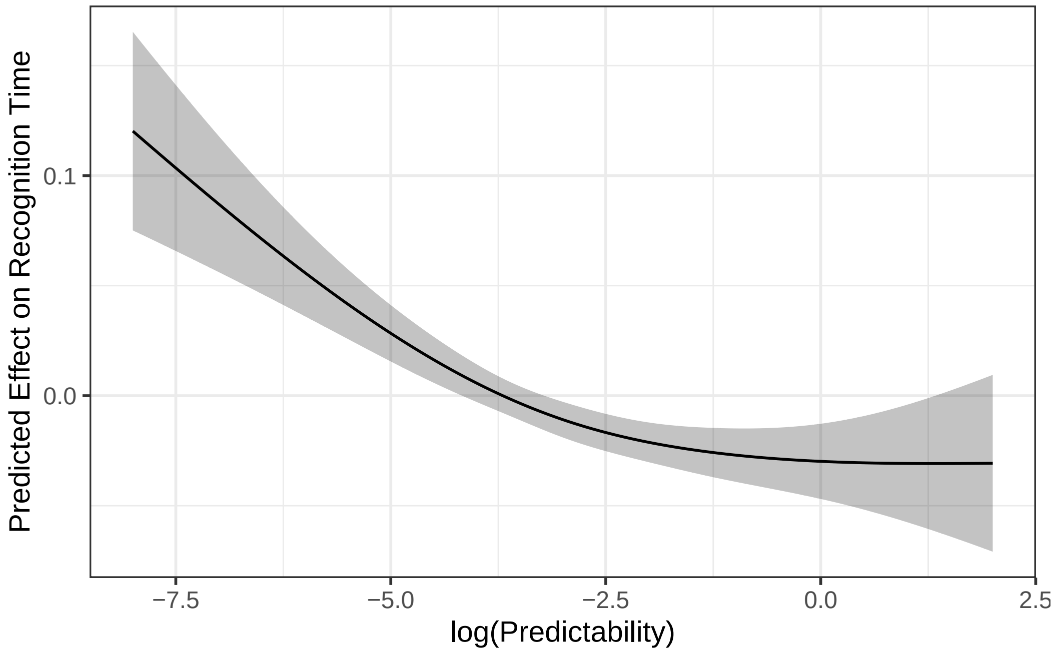


Figure 3.3.4.: Visualization of the model results from Table 3.3.4 for frequency (top) and predictability (bottom). Frequencies are per million.

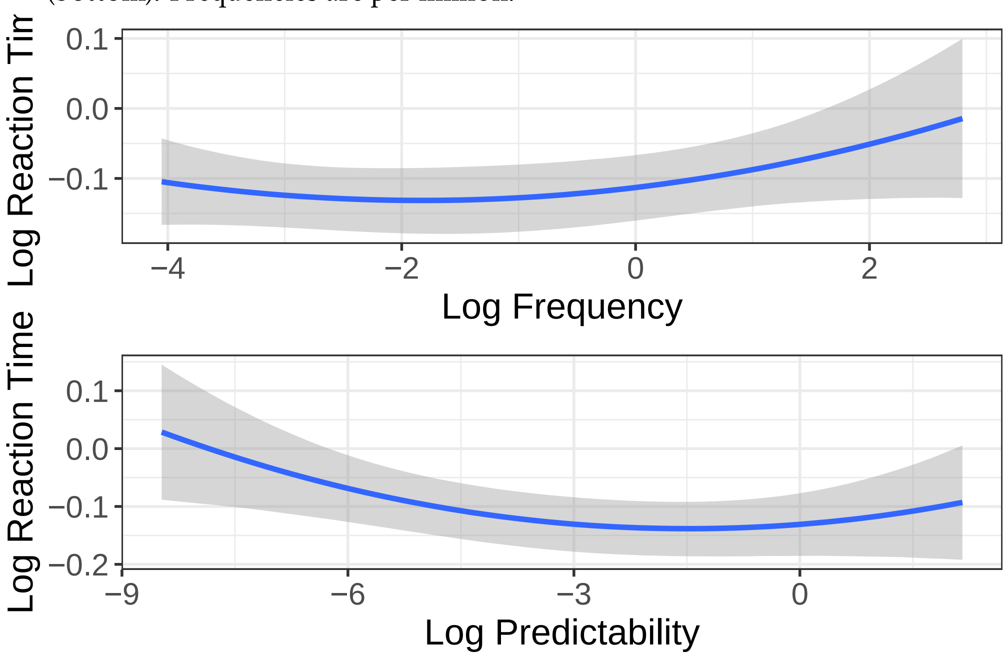
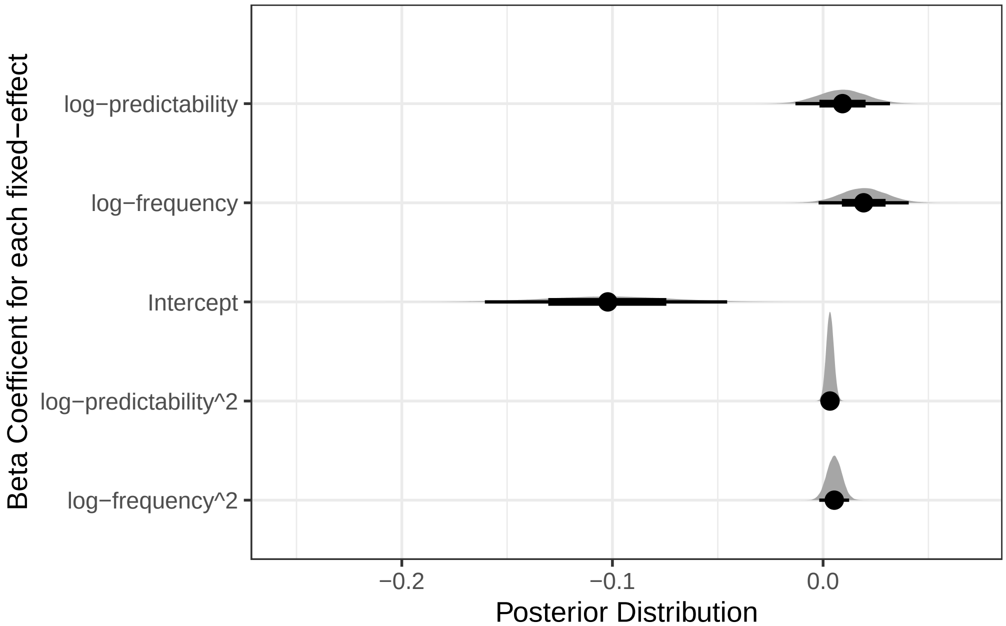


Figure 3.3.5.: Plot of the posterior distribution for the beta value of each fixed-effect in our Bayesian quadratic regression model. The y-axis contains the different fixed-effects and the x-axis contains the posterior distribution of beta values for the corresponding fixed-effect.



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* (e.g., *pick the ball up*) constructions and *Verb-Particle-Object* (e.g., *pick up the ball*) constructions are two distinct constructions,[8] as opposed to being two alternative realizations of a single constructions. In contrast, non-phrasal verbs can be generated through compositional knowledge (e.g., *walk up*). This suggests that phrasal verbs may be stored holistically regardless of frequency/predictability, while non-phrasal verbs may be generated compositionally unless they are frequent or predictable enough. 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

[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.

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. 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.

Chapter 4.

Emergent Ordering Preferences in Large Language Models

4.1. Introduction

Large language models have stormed the media in the last few years, becoming a popular topic in the scientific literature. Their rise to fame has brought with them many heated debates regarding whether large language models constitute human-like models of language or whether their behavior is completely different from humans (Bender et al., 2021; Bender & Koller, 2020; Piantadosi, 2023; Piantadosi & Hill, 2022).

Many of these debates have centered around the tradeoff between computation and storage: how much are these models simply reproducing from their training data vs how much of their productions are novel utterances using learned linguistic patterns? On one hand, there is no doubt that large language models store and reproduce large chunks of language. In fact, OpenAI is even being sued by *The New York Times* for allegedly reproducing entire articles verbatim.[1] This sentiment – that large language models are nothing but glorified copy cats – has been echoed by several other prominent linguists (Bender et al., 2021; Bender & Koller, 2020; c.f., Piantadosi & Hill, 2022).

Specifically, proponents of the "LLMs as copy cats" argument have pointed out that large language models are trained on an inconceivably large amount of data. For example, the OLMo models

[1] https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

were trained on trillions of tokens (Groeneveld et al., 2024).[2] As such, it is difficult to determine how much of the text generated by an LLM is truly novel, and how much is simply reproduced from its training data. This is further complicated by the fact that training data for LLMs is typically either not publicly available, or so huge that it's incredibly difficult to work with. On the other hand, it is clear that large language models are learning at least some linguistic patterns. For example, McCoy et al. (2023) demonstrated that GPT-2 is able to generate well-formed novel words as well as well-formed novel syntactic structures, despite copying extensively.

A similar debate in the field has centered around whether large language models learn any knowledge about the meaning of words. For example, Bender & Koller (2020) have argued that large language models, which are only trained on the form, have no way of learning anything about meaning. They pointed out that large language models do not have the rich information that humans receive, such as the referent of the form. However, Piantadosi & Hill (2022) rebutted this claim by arguing that co-occurrence statistics can be extremely informative about a word's meaning. For example, they argued that many words, such as "justice", contain no clear referent and instead have to be learned by humans based on the context that they occur in. It seems plausible that large language models could learn at least some information about the meaning of words in a similar manner.

These debates, however, have been highly theoretical and speculative and very few empirical studies have been done to actually investigate these questions LeBrun et al. (2022). Thus in the present paper we address these debates by taking an in-depth look at large language models' abilities to learn abstract knowledge beyond simply the statistics of individual tokens. Specifically we examine the ordering of novel binomials in English (Noun *and* Noun constructions, e.g., *cats and dogs*). The nouns in binomials can be ordered without affecting the meaning of the phrase much (e.g., *cats and dogs* vs *dogs and cats*). Despite this, human preferences for which noun should be placed first vary in strength. For example, there is a pretty strong preference for *bread and butter* as opposed to *butter and bread*. However, both *computers and monitors* and *monitors and computers* are natural. Binomials are a

[2]This is magnitudes larger than the 350 million words that the average college-aged speaker has seen in their lifetime (Levy et al., 2012).

useful test case because there is a great deal of evidence demonstrating that human ordering preferences are driven by abstract preferences, such as a preference for more powerful words to be placed first (e.g., *god and man* vs *man and god*). Thus by examining binomials varying in frequency, we can gain insight into whether large language models are learning abstract preferences and to what degree they learn similarly to humans.

4.1.1. Abstractions in Large Language Models

The evidence for learned abstractions in large language models is extremely mixed. Investigations into BERT have yielded mixed results for their ability to learn and apply abstract knowledge (Haley, 2020; Lasri et al., 2022; Li et al., 2023; Li & Wisniewski, 2021). For example, Haley (2020) demonstrated that many of the BERT models are not able to reliably determine the plurality of novel words. Additionally, Li & Wisniewski (2021) demonstrated that when tasked with producing the correct tense for a word, BERT tends to rely on memorization from its training data as opposed to learning the more general linguistic pattern.

On the other hand, Lasri et al. (2022) demonstrated that BERT can generalize well to novel subject-verb pairs. They tested BERT's performance on novel sentences along with semantically incoherent but syntactically sensible sentences (e.g., *colorless green ideas sleep furiously*) and found that when masking the verb for these sentences BERT still applies higher probability to the correct inflection of the verb. Additionally, Li et al. (2023) demonstrated that BERT is able to use abstract knowledge to correctly predict subject-verb and object-past participle agreements in French.

Research using other language models have also yielded similar results. For example, as mentioned earlier, McCoy et al. (2023) found that while GPT-2 copies extensively, it also produces both novel words as well as novel syntactic structures. Additionally Misra & Mahowald (2024). They examined whether a language model trained on a comparable amount of data as humans can learn article-adjective-numeral-noun expressions (a beautiful five days). Specifically, without having a great deal of experience with them, humans learn that *a beautiful five days* is perfectly grammatical, but *a blue* *five days* is not. Misra & Mahowald (2024) demonstrated that language models learn this even if they have no AANNs in their training data. They further demonstrated that they do this by generalizing across similar constructions, such as *a few days*. Further, Yao et al. (2025) examined whether language models trained on a comparable amount to humans can learn the length and animacy preferences that drive dative alternations in humans. They found that language models can learn these preferences, even without much experience with the dative alternation. These results suggest that language models can learn generalizations without a great amount of data.

There is also evidence that transformer models can learn abstractions from other domains as well. For example, Tartaglini et al. (2023) examined the ability of a transformer model in a same-different task (i.e., determining if two entities in an image are the same). They found that some models can reach near perfect accuracy on items they have never seen before.

Finally, there's evidence that inducing abstractions facilitates performance in large language models. For example, Zheng et al. (n.d.) used a novel prompting technique to enable LLMs to use abstractions when reasoning. They found that LLMs hallucinate less when they implement abstractions in their reasoning. Similarly, McCoy et al. (n.d.) demonstrated that large language models can use abstractions, such as an abstract preference for certain syllable structures, to learn language more easily. Their results suggest that inducing abstractions may help reduce the amount of training that large language models require.

4.1.2. Abstractions in Humans

Abstractions have been a part of just about every linguistic theory, including both generativist and non-generativist theories. This is not surprising since one of the hallmarks of human language learning is the ability to produce novel, never-heard-before utterances. In order to do so, most theories posit that humans leverage our remarkable ability to learn linguistic patterns beyond simple co-occurrence rates (c.f., Ambridge, 2020). For example, when presented a novel noun, children are able to consistently produce the proper plural form of that noun (Berko, 1958). Similarly, children are able to leverage similarities across different contexts to learn a word's general meaning (Yu & Smith, 2007).

There have been several studies showing human ordering preferences for binomials are driven, at least in part, by abstract ordering preferences (Morgan & Levy, 2015, 2016a, 2024). In order to capture the abstract ordering preferences of humans across binomial constructions, Morgan & Levy (2016a) developed a model to quantify the abstract ordering preference of a given binomial in English. They demonstrated that the model's predicted abstract ordering preferences are not the same as the observed preferences in corpus data. They further demonstrated that human ordering preferences for low-frequency items are primarily driven by abstract ordering preferences, and their preferences for high-frequency items are driven primarily by the observed preferences in corpus data. They operationalized frequency using the overall frequency of a binomial, i.e. the total frequency in both possible orders. This provides a measure of expression frequency that is not confounded with the frequency of a specific order.

Since human ordering preferences deviate from the observed preferences [i.e., humans aren't simply reproducing binomials in the same order they they heard them; Morgan & Levy (2024)], ordering preferences thus present a useful test case for large language models. If large language models learn representations beyond simply memorizing the training dataset or superficially reproducing word co-occurrences, they may learn abstract ordering preferences similar to humans, and this may be reflected in their binomial ordering preferences.

4.1.3. Present Study

In the present study we examine whether large language models are simply copying their input, or whether they are behaving more similarly to humans and learning abstract linguistics patterns. We use binomials as a test case because human ordering preferences deviate from the observed preferences for them. While binomials are a single linguistic construction, they are well-studied in the linguistics literature and thus provide us with a strong human baselines that we can compare the LLM's performance to.

In Experiment 1, we examine whether large language models are sensitive to ordering preferences for binomials that range from low-frequency to high-frequency. In Experiment 2, we go more in-depth and explore whether OLMo-7B (Groeneveld et al., 2024) is sensitive to abstract ordering preferences for novel binomials that the model has never seen before. Finally, in Experiment 3, we examine the same questions at different stages of the model's training in order to determine how these abstract ordering preferences emerge as a function of the training.

4.2. Experiment 1

4.2.1. Methods

Dataset

In order to examine the ordering preferences of binomial constructions in large language models, we use a corpus of binomials from Morgan & Levy (2015). The corpus contains 594 binomial expressions which have been annotated for various phonological, semantic, and lexical constraints that are known to affect binomial ordering preferences. The corpus also includes:

1. The estimated generative preferences for each binomial representing the ordering preference for the alphabetical ordering (a relatively unbiased reference form), estimated from the above constraints (independent of frequency). The generative preferences take a value between 0 and 1, with 0 being a stronger preference for the nonalphabetical form, and 1 being a stronger preference for the alphabetical form. The generative constraints were calculated using (Morgan & Levy, 2015)'s model.

2. The observed binomial orderings which are the proportion of binomial orderings that are in alphabetical order for a given binomial, gathered from the Google *n*-grams corpus (Lin et al.,

2012). The Google *n*-grams corpus is magnitudes larger than the language experience of an individual speaker and thus provides reliable frequency estimates.

3. The overall frequency of a binomial expression (the number of times the binomial occurs in either alphabetical or non-alphabetical order). Overall frequencies were also obtained from the Google *n*-grams corpus (Lin et al., 2012).

Language Model Predictions

In order to derive predictions for large language models, we used the following models from the GPT-2 (Radford et al., 2019) family, the Llama-2 (Touvron et al., 2023) family, Llama-3 family (https://github.com/meta-llama/llama3), and the OLMo (Groeneveld et al., 2024) family. From smallest to largest in number of parameters: GPT-2 (124M paramters), OLMo 1B (1B parameters), GPT-2 XL (1.5B parameters), Llama-2 7B (7B parameters), OLMo 7B (7B parameters), Llama-3 8B (8B parameters), Llama-2 13B (13B parameters), and Llama-3 70B (70B parameters). For each model, we calculated the ordering preferences of the alphabetical form for each binomial in the dataset. The predicted probability of the alphabetical form was calculated as the product of the model's predicted probability of each word in the binomial. In order to accurately calculate the probability of the first word in the binomial, each binomial was prepended with the prefix "Next item:". Thus the probability of the alphabetical form, *A and B* is:

$$\begin{aligned} P_{alphabetical} &= P(A|Next\,item:) \\ &\times P(and|Next\,item:A) \\ &\times P(B|Next\,item:A\,and) \end{aligned}$$

where *A* is the alphabetically first word in the binomial and *B* is the other word. Additionally, the probability of the nonalphabetical form, *B and A* is:

$$
\begin{aligned}
P_{nonalphabetical} = & P(B|Next\, item:) \\
& \times P(and|Next\, item: B) \\
& \times P(A|Next\, item: B\, and)
\end{aligned}
$$

Finally, to get an overall ordering preference for the alphabetical form, we calculated the (log) odds ratio of the probability of the alphabetical form to the probability of the nonalphabetical form:

$$LogOdds(AandB) = log(\frac{P_{alphabetical}}{P_{nonalphabetical}})$$

Analysis

The data was analyzed using Bayesian linear regression models, implemented in *brms* (Bürkner, 2017) with weak, uninformative priors. For each model, the dependent variable was the log odds of the alphabetical form to the nonalphabetical form. The fixed-effects were abstract ordering preference (represented as *AbsPref* below), observed preference (*ObservedPref*), overall frequency (*Freq*), an interaction between overall frequency and abstract ordering preference (*Freq:AbsPref*), and an interaction between overall frequency and observed preference (*Freq:ObservedPref*). The model equation is presented below:

$$
\begin{aligned}
LogOdds(AandB) \sim & AbsPref \\
& + ObservedPref \\
& + Freq \\
& + Freq : AbsPref \\
& + Freq : ObservedPref
\end{aligned}
\tag{4.1}
$$

Frequency was logged and centered, and abstract ordering preference and observed preference were centered such that they ranged from -0.5 to 0.5 (instead of from 0 to 1). Note that since abstract ordering preference and observed preference are on the same scale, we can directly draw comparisons between the coefficient estimates for these fixed-effects in our regression model.

4.2.2. Results

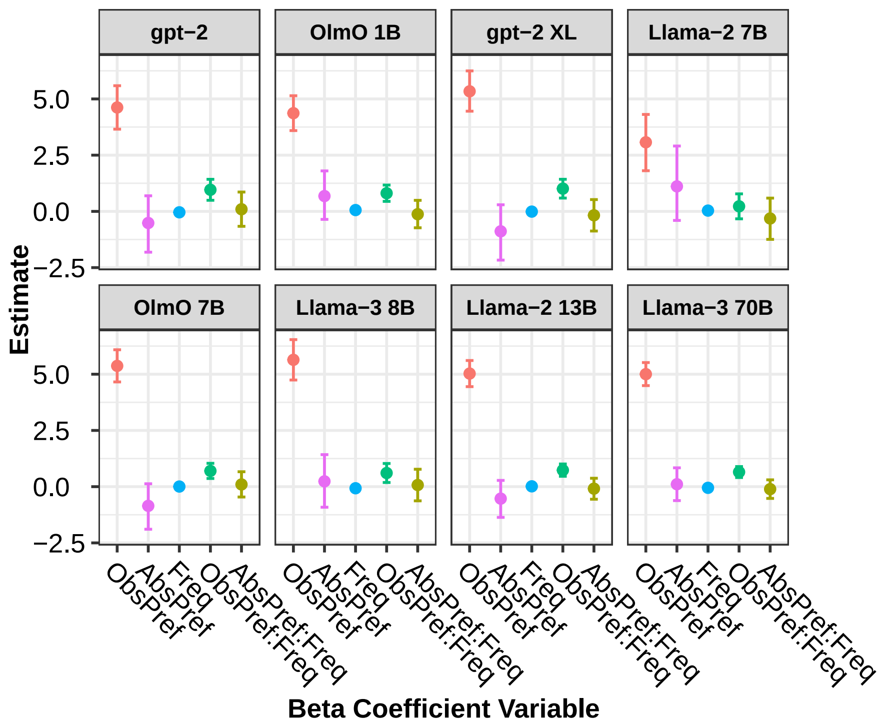
Our full model results are presented in the appendix (Table A.0.1) and visualized in Figure 4.2.1. For each model, the figure shows the values for each of the coefficients from the model in Equation 4.1, representing how strongly each language model relies on observed preference, abstract ordering preference, overall frequency, the interaction between abstract ordering preference and overall frequency, and the interaction between observed preference and overall frequency.

Our results are similar across all the large language models we tested. Specifically, we find no effect of abstract ordering preferences and no interaction effect between abstract ordering preference and overall frequency. We do find an effect of observed preference suggesting that the models are mostly reproducing the ordering preferences found in their training. We also find an interaction effect between observed preference and overall frequency, suggesting that the effect of observed frequency is stronger for high-frequency items.

4.2.3. Discussion

In the present study we examined the extent to which abstract ordering preferences and observed preferences drive binomial ordering preferences in large language models. We find that their ordering preferences are driven primarily by the observed preferences. Further, they rely more on observed preferences for higher frequency items than lower frequency items. Finally, they don't seem to be using abstract ordering preferences at all in their ordering of binomials.

Figure 4.2.1.: Results for each beta coefficient estimate from each model. Models are arranged from smallest to largest from left to right. The x-axis contains each coefficient and the y-axis contains the predicted beta coefficient of the respective model. Error bars indicate 95\% credible intervals.



Our results give us insight into the differences between humans and large language models with respect to the ways in which they trade off between abstract and observed preferences. For example, our dataset contains low-frequency binomials (e.g., *alibis and excuses*), including binomials that a college-age speaker would have heard only once in their life. Due to their low frequency, humans rely substantially on abstract ordering preferences to process these lower frequency items (Morgan & Levy, 2024). This is not the case, however, for large language models, which rely exclusively on observed preferences for these items. This is true even for the smallest models we tested, such as GPT-2. We conclude that, although large language models can produce human-like language, they accomplish this in a quantitatively different way than humans do: they rely on observed statistics from the input in at least some cases when humans would rely on abstract representations.

4.3. Experiment 2

In Experiment 1, we demonstrated that large language models don't use abstract ordering preferences when producing binomials. However, it is possible that this is because they have experienced the binomials before, even the low-frequency ones. Thus in Experiment 2 and Experiment 3 we examine whether OLMo-7B is sensitive to abstract ordering preferences for novel binomials that the model has never seen before. We also examine the individual constraints that drive abstract ordering preferences in humans, such as the preference for short words before long words, to determine whether OLMo is sensitive to the same constraints in the same way as humans.

Specifically, in Experiment 2, we examine whether OLMo-7B's ordering preferences are driven by abstract ordering preferences for novel binomials. In order to do so, we created a list of binomials and searched the Dolma corpus we created to confirm that they did not occur in either alphabetical or nonalphabetical ordering. We then coded the binomials for each of the constraints mentioned earlier. Finally, we examined whether OLMo-7B shows any preference for one ordering over the other for each binomial. If OLMo has developed any abstract ordering preferences, it should show a systematic preference for one ordering over the other. If it has not, then it should show no preference for one ordering over the other.

4.3.1. Datasets

Dolma

For Experiments 2 and 3, we use the dataset described in this section. In order to examine whether large language models learn preferences above and beyond simply memorizing co-occurrence rates, we created a 1-grams, 2-grams, and 3-grams corpus of Dolma (Soldaini et al., 2024). Specifically, we used Dolma version 1_7 (2.05 trillion tokens), which was used to train OLMo-7B-v1.7 (Groeneveld et al., 2024). Our corpus contains every n-gram (ignoring punctuation and capitalization) in the Dolma corpus, as well as the number of times that n-gram appeared.

We then created a list of binomials and searched the corpus to find a list of binomials that did not occur. We eliminated binomials which occurred more than zero times in either possible ordering. Thus, OLMo has had no experience with either ordering of any of our binomials. We also calculated their unigram and bigram frequencies for each binomial. Our full list of items comprises 131 binomials and is reproduced in the appendix section (Section Appendix C).

Abstract Ordering Preferences Corpus

In order to examine whether large language models are learning preferences similar to humans, we calculated the abstract ordering preference value for each of our binomials (following Morgan & Levy, 2016a). Morgan & Levy (2016a) demonstrated that their model's estimated abstract ordering preference value is a significant predictor of human binomial ordering preferences, even after accounting for the frequency of each ordering. Abstract ordering preferences are calculated from a mix of semantic and phonological properties that human binomial ordering preferences have been shown to be sensitive to (Benor & Levy, 2006). For each of these constraints, a positive value indicates a preference for the alphabetically first word to be placed first (a neutral reference order). A negative value indicates a preference for the nonalphabetical word to be placed first. For example, a positive value of frequency indicates that the alphabetical word is more frequent and thus is predicted to be placed first, while a negative value indicates that the nonalphabetical word is more frequent. The constraints along with the estimated weights in humans are as follows (taken from Morgan & Levy, 2015):

- **Length**: The shorter word should appear first, e.g. *abused and neglected*. In human data, the weight of this constraint was estimated to be 0.15.
- **No Final Stress**: The final syllable of the second word should not be stressed, e.g. *abused and neglected*. In human data, the weight of this constraint was estimated to be 0.36.
- **Lapse:** Avoid unstressed syllables in a row, e.g. *FARMS and HAY-fields* vs *HAY-fields and FARMS.* In human data, the weight of this constraint was estimated to be 0.19.
- **Frequency**: The more frequent word comes first, e.g. *bride and groom*. In human data, the weight of this constraint was estimated to be 0.09.
- **Formal Markedness**: The word with more general meaning or broader distribution comes first, e.g. *boards and two-by-fours*. In human data, the weight of this constraint was estimated to be 0.24.
- **Perceptual Markedness**: Elements that are more closely connected to the speaker come first. This constraint encompasses Cooper & Ross (1975)'s 'Me First' constraint and includes numerous subconstraints, e.g.: animates precede inanimates; concrete words precede abstract words; e.g. *deer and trees*. In human data, the weight of this constraint was estimated to be 0.25.
- **Power**: The more powerful or culturally prioritized word comes first, e.g. *clergymen and parishioners*. In human data, the weight of this constraint was estimated to be 0.26.
- **Iconic/scalar sequencing**: Elements that exist in sequence should be ordered in sequence, e.g. *achieved and maintained*. In human data, the weight of this constraint was estimated to be 1.30.

- **Cultural Centrality**: The more culturally central or common element should come first, e.g. *oranges and grapefruits*. In human data, the weight of this constraint was estimated to be 0.42.
- **Intensity**: The element with more intensity appear first, e.g. *war and peace*. In human data, the weight of this constraint was estimated to be 0.02.

Morgan & Levy (2015) then used a logistic regression model to combine these constraints into a single constraint (overall abstract preference) for each binomial.

4.3.2. Methods

Language Model Predictions

We use the same methods as Experiment 1 to obtain language model predictions for our items.

Analyses

We present three Bayesian linear mixed-effects models implemented in *brms* (Bürkner, 2017) with weak, uninformative priors. For each of our models, the intercept represents the grand mean and the coefficient estimates represent the distance of the effect from the grand mean. Bayesian statistics don't force us into a binary interpretation of significance, however we can consider an estimate to be statistically significant if the credible interval for that estimate excludes zero.

For all three analyses, the dependent variable is LogOdds(AandB), which was described above. Our dependent variable in the first analysis is the abstract ordering preference for each binomial (AbsPref). In order to rule out bigram probabilities as driving the model's preferences, our second analysis contains the binomial's bigram probabilities (the odds ratio of product of bigram probabilities of the alphabetical order to the product of bigram probability of the nonalphabetical order) as well. Finally, our dependent variables in the third analysis are the individual constraints that are used to calculate AbsPref. The model equations are below in Equation 4.2, Equation 4.3, and Equation 4.4. Note that Formal Markedness and Iconicity were dropped from the second model because the constraint values were zero for all of the binomials. Further, our constraints demonstrated a level of co-linearity. Co-linearity can result in poor model estimates and inflated credible intervals. In order to deal with this, we dropped the constraint with the highest variance inflation factor (which turned out to be the lapse constraint). All other constraints had a variance inflation factor value below 1.5. We then performed backward model selection and dropped the predictors whose credible intervals were most centered around zero until the remaining predictors' credible intervals were all at least 75% greater than or less than zero. This resulted in dropping the no final stress, intense, and percept constraints. We acknowledge that this approach is quite exploratory and thus interpretations at the level of the individual constraint must be taken with a grain of salt.

$$LogOdds(AandB) \sim AbsPref \tag{4.2}$$

$$LogOdds(AandB)\ AbsPref \cdot BigramProbs \tag{4.3}$$

$$LogOdds(AandB) \sim Culture + Power + Freq + Len \tag{4.4}$$

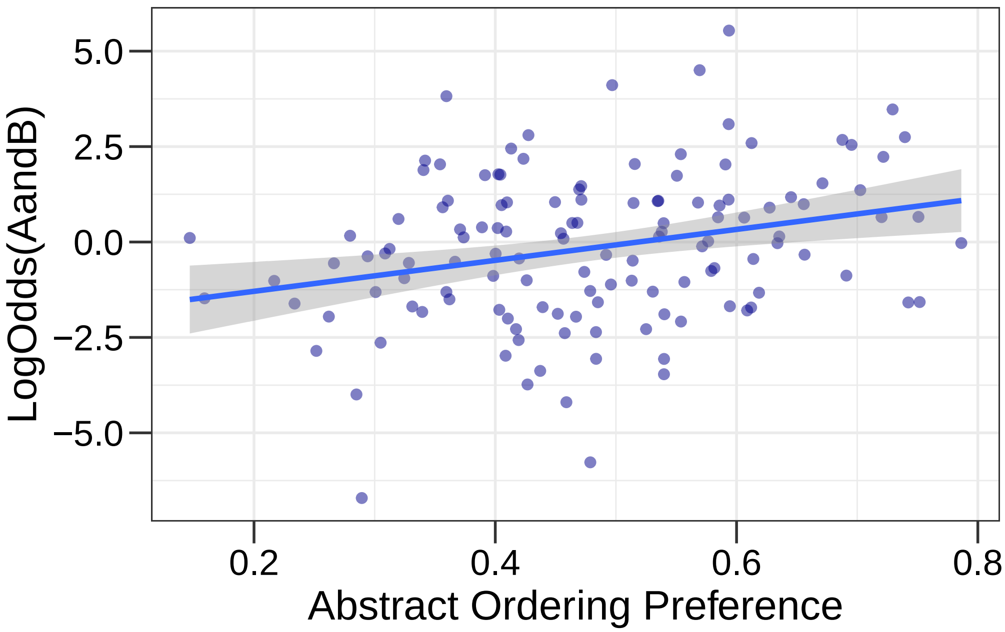
4.3.3. Results

The results for the first analysis are presented below in Table 4.3.1. Our results suggest that there is a main-effect of abstract ordering preference for OLMo's 7B model. A visualization of these results can be found below in Figure 4.3.1.

Table 4.3.1.: Model results examining the effect of AbsPref on LogOdds(AandB).

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	-1.370	0.636	-2.672	-0.222
AbsPref	2.547	1.274	0.291	5.151

Figure 4.3.1.: Visualization of the effects of AbsPref on LogOdds(AandB)



The results of our second model suggest that the model's ordering preferences were not driven by bigram probabilities (see Table 4.3.2).

Table 4.3.2.: Model results examining the effect of AbsPref and Bigram Probabilities on LogOdds(AandB).

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	-1.224	0.930	-3.299	0.289
AbsPref	2.444	1.852	-0.402	6.631
BigramProbs	0.495	0.747	-0.915	2.002
AbsPref:BigramProbs	0.168	1.157	-2.126	2.601

Table 4.3.2.: Model results examining the effect of AbsPref and Bigram Probabilities on LogOdds(AandB).

	Estimate	Est.Error	Q2.5	Q97.5

While these results suggest that the large language models' ordering preferences are sensitive abstract ordering preferences and not bigram probabilities, it's unclear whether their behavior is similar to humans on the level of the individual constraints. Thus, in the third analysis we examined which specific constraints the model is sensitive to, and to what extent.[3] For this analysis, following Houghton et al. (2024), we also present the percentage of posterior samples greater than zero. The results of this analysis can be found below in Table 4.3.3.

Table 4.3.3.: Model results examining the effect of each individual constraint on LogOdds(AandB).

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-0.132	0.163	-0.451	0.185	20.750
Culture	0.414	0.254	-0.078	0.916	94.945
Power	0.720	0.263	0.204	1.236	99.665
Freq	0.091	0.087	-0.079	0.263	85.170
Len	-0.209	0.134	-0.476	0.052	5.870

The model is most sensitive to the Power constraint, however there appears to be a marginal effect of Culture as well, since nearly 95% of the posterior samples are greater than zero despite the credible interval crossing zero. Surprisingly, there also appears to be a negative effect of length with a slight preference to place the longer word first, which is the opposite direction from what we see in humans. Length is often correlated with frequency, since frequent words tend to be shorter. As

[3] It's also important to consider how many of our binomials the constraint even applied to (i.e., how many binomials were the constraints non-zero). For the Culture constraint, 62 of our 131 binomials had a non-zero value. For the Power constraint, 54 were non-zero, for the Frequency constraint, all 131 binomials were non-zero, and for the Len constraint, 85 were.

such, we ran a model without frequency to determine whether the negative effect of length was do to co-linearity with frequency. However, dropping frequency from the model did not affect the effect of length. Further, we also ran a model with only length as the predictor and for that model as well the estimate of length remained negative.

4.3.4. Discussion

Experiment 2 found that OLMo-7B has learned abstract ordering preferences even for novel binomials that it has never seen before. Further, these ordering preferences aren't simply based on the individual word or bigram frequencies. Specifically, we find a main-effect of abstract ordering preferences on the model's binomial ordering preferences. Additionally, we find a strong preference to place the more powerful word first, a weak preference to place the more culturally central word first, and a weak preference to place the longer word first.

These results together suggest that the model is learning abstract ordering preferences but in a way that is not identical to humans. For example, while both LLMs and humans show a preference for placing the more powerful words first and the more culturally central word first, humans also show a sensitivity to formal markedness, perceptual markedness, and frequency (Morgan & Levy, 2016a), which we do not find evidence for in large language models' binomial ordering preferences. Additionally, humans prefer to place the shorter word first Morgan & Levy (2015). However, we find the opposite finding here: large language models prefer to place the longer word first. One explanation for this is a difference in terms of the input between humans and large language models. The length constraint is determined by the number of syllables. While syllables are salient cues in the audio that humans receive during learning, it's less clear how salient of a cue they are for large language models which receive sub-word tokens (which vary in their size, from being individual orthographic symbols to being entire words.

However, it's unclear how large language models learn these preferences in the first place. Thus, in Experiment 3 we examine these constraints at different points in OLMo's training.

4.4. Experiment 3

In Experiment 2 we demonstrated that large language models are not simply copying their training, but are learning some abstract ordering preferences from their input. However, OLMo makes public various checkpoints during the model's training, thus allowing us the opportunity to examine how these preferences arise as a function of the training. Thus, in Experiment 3 we examine the evolution of these learned abstract ordering preferences as the model learns over time.

4.4.1. Methods

Language Model Predictions

The language model predictions in Experiment 3 were obtained using the same procedure as in Experiments 1 and 2. However, instead of calculating these metrics only for the main model, we calculated them at various checkpoints. These checkpoints are listed below, in terms of the steps as well as the number of billions of tokens the model had been trained on at that checkpoint:

- Step 0, 0B Tokens
- Step 1000, 2B Tokens
- Step 10000, 41B Tokens
- Step 50000, 209B Tokens
- Step 100000, 419B Tokens
- Step 200000, 838B Tokens
- Step 400000, 1677B Tokens

Analysis

The analyses in Experiment 3 were identical to Experiment 2, however we ran these analyses for each of the checkpoints listed above.

4.4.2. Results

Our model estimates for the effect of AbsPref on LogOdds(AandB) at each checkpoint are presented below in Table 4.4.1 and visualized in Figure 4.4.1.

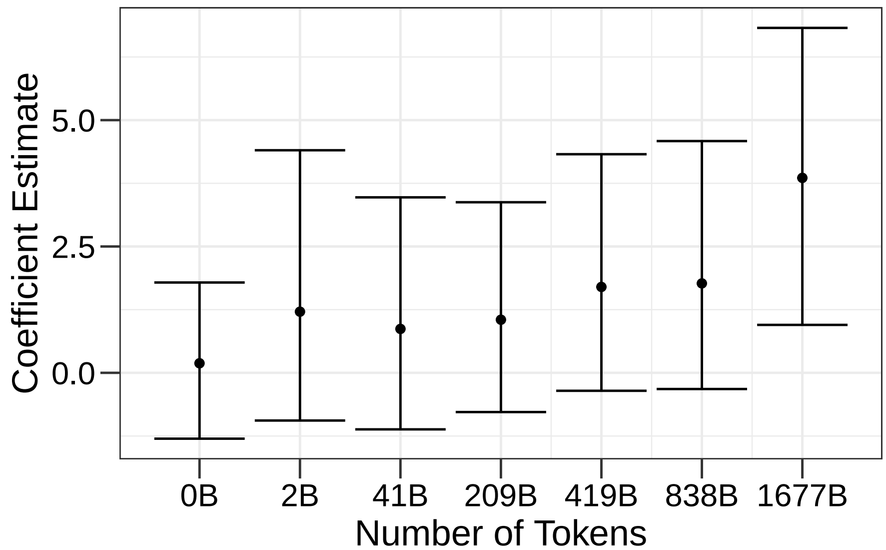
Table 4.4.1.: Model results examining the effect of AbsPref on LogOdds(AandB) for each checkpoint.

Number of Tokens	Estimate	Est.Error	Q2.5	Q97.5
0B	0.189	0.789	-1.303	1.787
2B	1.210	1.346	-0.944	4.404
41B	0.869	1.167	-1.119	3.472
209B	1.051	1.040	-0.776	3.375
419B	1.700	1.216	-0.355	4.326
838B	1.770	1.267	-0.320	4.585
1677B	3.858	1.516	0.948	6.825

The model results are visualized below in Figure 4.4.1.

Our results demonstrate that it takes quite a large number of tokens for the model to learn the abstract ordering preferences. As Figure 4.4.1 demonstrates, the effect of abstract ordering preference isn't convincing until the model has experienced 1677B tokens. However, it does appear that the model develops a slight preference quite rapidly. For example, by 2 billion tokens there appears to be a very slight (though unconvincing) effect of abstract ordering preferences on the ordering of binomials.

Figure 4.4.1.: Visualization of the model predictions for the effect of AbsPref on LogOdds(AandB) for each checkpoint.



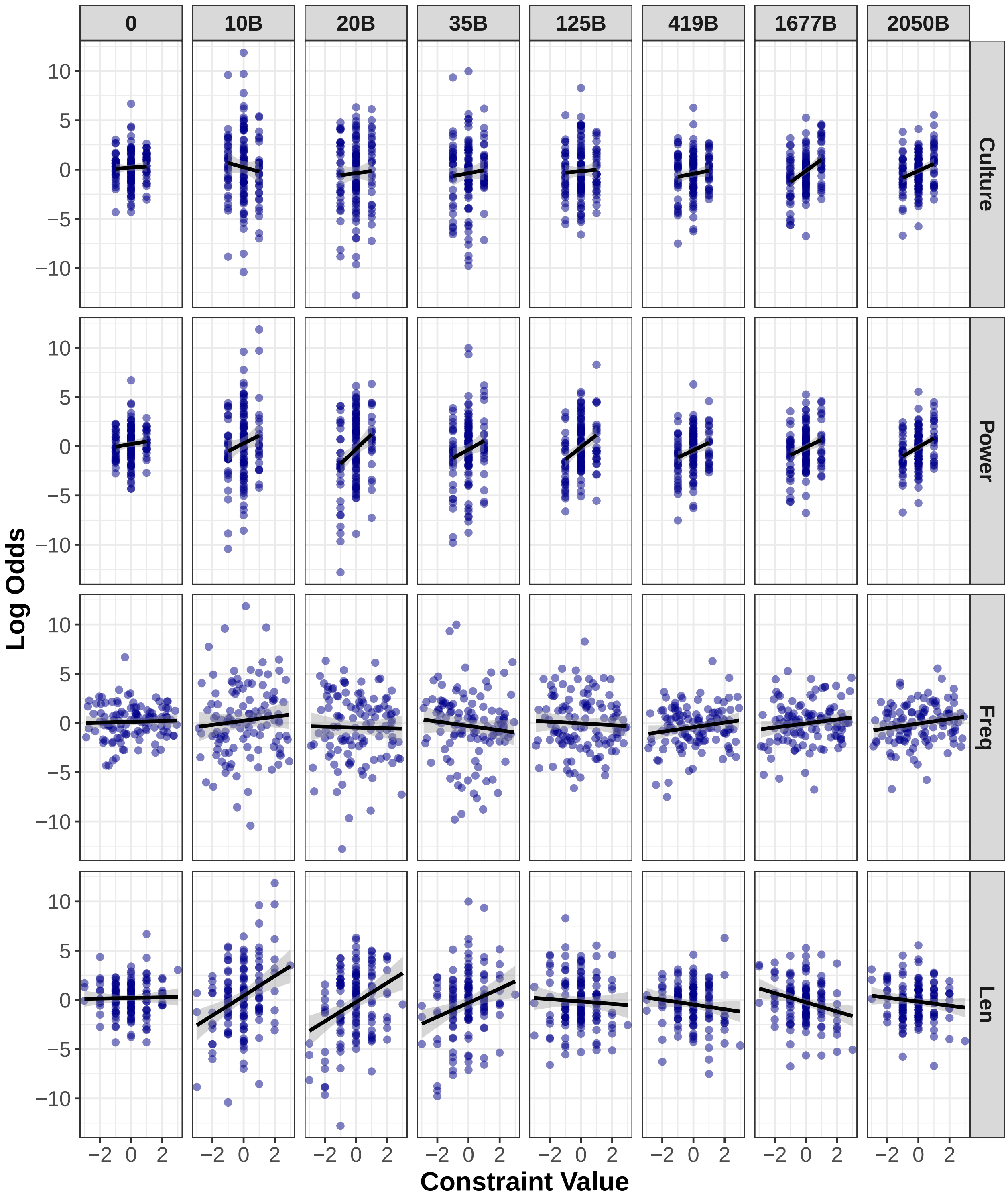
Similar to Experiment 2, in our second analysis we present a breakdown of the effects of each individual constraint. In this analysis, however, we demonstrate the effect of each constraint at each checkpoint. The full table results can be found in the the appendix section (Appendix B), but we present a visualization below in Figure 4.4.2.

Interestingly, it appears that early on the model already shows evidence of learning human-like preferences. For example, by 10 billion tokens the model has learned to place more powerful words first and shorter words first. However, the model seems slower to learn to place more culturally central words first. Further, as it receives more training the effect of length undergoes a reversal in direction.

4.4.3. Discussion

Our results demonstrate that OLMo learns ordering preferences early on for the power, frequency, and length constraints, but is slower to learn ordering preferences for the culture constraint. Further, the model is human-like in its predictions for length early on, but as it receives more training data it learns the opposite length prediction. It is unclear what exactly is causing this reversal, but it

Figure 4.4.2.: Visualization of the effect of each constraint on the ordering preference at each checkpoint.



may be a function of the tokenization differences between human input and large language models' input.

It is interesting that the model is quick to learn the power constraint, but slower to learn the culture constraint. Both constraints require a level of world-knowledge but the model learns which entities are more powerful relatively quickly, but not which is more culturally central. One interesting question is whether humans also take longer to learn the culture constraint, or whether they learn this early on. If children learn this constraint early on it may suggest differences with respect to how easily large language models learn world knowledge. On the other hand, if children also take longer to learn the culture constraint it may suggest that the power constraint is simply easier to learn.

4.5. Conclusion

In the present study, we examined the ordering preferences of binomials in various large language models. We found that for binomials that the models have experienced, they do not use abstract preferences but instead reproduce them proportionately to their training data. This is the case even for low-frequency binomials. Interestingly, for novel binomials, we found that they do learn and use abstract ordering preferences.

Additionally, while overall the model does show evidence of having learned abstract preferences, these preferences are not identical to human ordering preferences. For example, the model shows a preference for longer words placed before shorter words, which is the opposite of what humans prefer.

Further, show that while the effect of abstract ordering preference on a whole takes a great deal of time (over 1677B tokens to be convincing), the model seems to learn human-like preferences at the level of some individual constraints quite early on.

Overall, our results suggest that large language models are not simply copying their input, but are learning interesting, human-like phenomena from their training. However, they are not learning identically to humans, as demonstrated by the opposite direction of the length preference. Further, while humans rely on abstract preferences even for binomials that they have encountered before, even high-frequency ones (Morgan & Levy, 2024), large language models rely on abstract ordering preferences only for items that they have not encountered at all. In other words, while large language models are able to learn abstract ordering preferences, in cases where humans would rely on these preferences, large language models seem to rely instead on their experience with the item.

Chapter 5.

Holistic Representations of Binomials in Large Language Models

5.1. Introduction

In the last few years large language models have surged in popularity and have remained in the center of both the media and the recent research. With their surge in popularity has come many debates about to what extent they constitute as effective models of human language (e.g., Bender et al., 2021; Piantadosi, 2023; Piantadosi & Hill, 2022). These questions have stemmed from clear differences in terms of both the training that the models receive as well as the performance of these models on language processing tasks. For example, common criticisms include their insanely large training size (sometimes being trained on upwards of 15 billion tokens), the potentially unrealistic nature of their tokenization (e.g., Chat GPT tokenizes kite as [*k*, *ite*][1]), and their poor performance on tasks that are trivial to humans (e.g., counting the number of *r*'s in *strawberry*[2]).

Many of these debates are centered around the extend to which large language models are actually learning something abstract from the data and to what extent they are simply regurgitating their training data. However, despite a large body of research, the results have been quite mixed with respect to the extent that they are learning something about the abstract linguistic structure as opposed to simply copying their training data (Haley, 2020; Lasri et al., 2022; Li et al., 2023; Li & Wisniewski, 2021;

[1] https://tiktokenizer.vercel.app/
[2] https://community.openai.com/t/incorrect-count-of-r-characters-in-the-word-strawberry/829618

McCoy et al., 2023; Misra & Mahowald, 2024; Yao et al., 2025). For example, Haley (2020) demonstrated that many BERT models are not able to reliably determine the correct plural form for novel words. Similarly, Li & Wisniewski (2021) demonstrated that BERT tends to rely on memorization from its training data when producing the correct tense of novel words.

In contrast, recent research has demonstrated BERT's ability to generalize well to novel subject-verb pairs (Lasri et al., 2022) and to use abstract knowledge to predict object-past participle agreements in French (Li et al., 2023). Further, McCoy et al. (2023) demonstrated that while GPT-2 copies extensively, it also produces both novel words as well as novel syntactic structures.

Additionally, in an attempt to address criticism about the unrealistic size of the training data for large language models, an interesting line of research has demonstrated that even smaller language models trained on an amount of data comparable to humans seem to be able to learn abstract linguistic knowledge from the data (Misra & Mahowald, 2024; Yao et al., 2025). For example, Misra & Mahowald (2024) examined whether a language model trained on a similar amount of data as humans could learn article-adjective-numeral-noun expressions (AANNs, e.g., *a beautiful five days*). They found that even after removing AANNs from the training data, language models are still able to learn which AANNs are appropriate and which ones are not (e.g., **a blue five days*). Additionally, Yao et al. (2025) examined whether a similar language model can learn the length and animacy preferences for the dative alternation (give X Y vs give Y to X). They found that a language model trained on a comparable amount of data as humans is able to learn these preferences. These results together demonstrate the ability for language models to learn general knowledge about the language.

Given the evidence that large language models show evidence of both learning abstract knowledge as well as copying extensively from their training data, it's unclear in what contexts they are leveraging their stored knowledge as opposed to leveraging their more abstract knowledge.

5.1.1. Stored Representations in Humans

There's a great deal of evidence that many multi-morphemic words and multi-word phrases are stored holistically (Bybee, 2003; Bybee & Scheibman, 1999; Stemberger & MacWhinney, 1986, 2004). For example, high-frequency phrases containing *don't* (e.g., *I don't know*) are more likely to be phonetically reduced than lower frequency words containing *don't* (Bybee & Scheibman, 1999). This suggests that higher-frequency phrases are represented holistically because the phonetic reduction cannot simply be attributed to phonetic reduction at the individual word-level.

Additionally, there is evidence from the psycholinguistics literature that high-frequency multi-word phrases have holistic representations (Morgan & Levy, 2015, 2016a, 2016b, 2024; O'Donnell, 2016; Siyanova-Chanturia et al., 2011). For example, Siyanova-Chanturia et al. (2011) demonstrated that binomials (e.g., *bread and butter*) are read faster in their frequent ordering than in their infrequent ordering. This suggests that reading times for multi-word phrases can't be reduced to the reading times of the individual words. However, it is possible that these faster reading times were driven by knowledge of generative preferences, such as a preference for short words before long words. Thus to investigate this, Morgan & Levy (2016a) examined whether human reading times for binomials is driven by generative preferences (e.g., a preference for more culturally central words first) or by experience with the binomial. Specifically, they examined whether human ordering preferences were driven simply by the relative frequency of the binomial. For example, *bread and butter* is vastly preferred over *butter and bread*. Is this driven by the fact that *bread and butter* is more frequent than *butter and bread* or driven by more abstract constraints, such as a preference to place the shorter word first? Interestingly, they found that high-frequency binomial ordering preferences are driven primarily by experience (i.e., the more frequent ordering is preferred) while low-frequency binomial ordering preferences are driven primarily by generative preferences (e.g., a preference for short words before long words). This suggests that humans are relying on generative knowledge for lower-frequency items, but relying on item-specific knowledge for high-frequency items, suggesting that high-frequency items (e.g., *bread and butter*) are stored holistically.

Given that humans rely on generative preferences for some binomials and item-specific preferences for other binomials, binomials present a good test case for examining this same trade off in large language models. Specifically, since humans holistically store high-frequency binomials holistically and rely more on generative knowledge for low-frequency binomials, if large language models are learning similarly, high-frequency binomials may be represented differently in each order, while compositional binomials may be represented similarly regardless of ordering.

5.1.2. Present Study

Since humans rely more on abstract knowledge for lower frequency items and rely more on their experience for high-frequency binomials (Morgan & Levy, 2024), a natural consequence of this is that they have learned separate representations for high-frequency binomials. If large language models are learning similarly to humans, they may also learn separate representations for high-frequency binomials but not for lower-frequency binomials.

The present study addresses this question by examining the semantic representations of binomials varying in relative frequency (the proportion of occurrences in one ordering to the other ordering) and overall frequency (the overall frequency of the binomial, regardless of ordering). We examine the embeddings for both ordering of binomials in a sentence context, as well as examine the embeddings for a compositional form of the binomial (which we will elaborate on in the methods section). We hypothesize that the representations of the more frequent form (higher relative frequency form) for binomials with a high overall frequency may diverge more from the compositional representation than the less frequent ordering (lower relative frequency form) does for the same binomial. That is, for high-frequency binomials, the representation for the more frequent ordering may be more different from the compositional representation than the less-frequent ordering. For lower-frequency binomials, large language models may not learn different representations for the different orderings of the same form, regardless of the relative frequency.

In Experiment 1, we examine the representations of different binomials across different large language models and in Experiment 2 we examine the timecourse of these representations across each hidden layer in OLMo's 1B model (Groeneveld et al., 2024).

5.2. Experiment 1

In Experiment 1 we examine the representations of binomials for GPT-2, GPT-2 XL (Radford et al., 2019), OLMo-1B, OLMo-7B (Groeneveld et al., 2024), and Llama2-7B (Touvron et al., 2023). We examine the representations for different binomials in sentence contexts as well as the compositional representations of those same binomials. We explain these metrics in detail below.

5.2.1. Methods

Dataset

Our dataset consists of 784 sentences containing binomials. The sentences have been annotated for both relative frequency and overall frequency. Relative frequency is operationalized as the proportion of occurrences in alphabetical order (a neutral reference order) to occurrences in nonalphabetical order. Overall frequency is operationalized as the count of *A and B* plus the count of *B and A*. Counts were obtained using the Google *n*-grams corpus (Lin et al., 2012).

Semantic Embeddings

In order to examine the semantic compositionality of binomials, we examined the semantic embeddings of five different large language models: GPT-2, GPT-2 XL (Radford et al., 2019), Llama-2 7B (Touvron et al., 2023), OLMo 1B and OLMo 7B (Groeneveld et al., 2024).[3]

For each LLM we examined the semantic embeddings of the binomials in a sentence context. We accomplished this by passing the sentence through each large language model and extracting the

[3] All of our code can be found publicly available at https://github.com/znhoughton/LLM-Storage.

second-to-last hidden layer for each of the words in the binomial. Since LLMs generate an embedding for each word, we computed the mean of these embeddings to represent the semantic embedding of the entire binomial in a sentence context (hereafter referred to as holistic embeddings). Next, we obtained the embedding for each word in the binomial individually, outside of a sentence context. We then computed the mean of these embeddings to represent the semantic embedding of the compositional form of the binomial (hereafter referred to as the compositional embeddings).

We then measured the cosine similarity between the holistic embeddings and the compositional embeddings for the alphabetical and nonalphabetical forms of each binomial. This is presented mathematically in Equation 5.1 and Equation 5.2, where cos_{α} is the cosine similarity between the holistic embeddings of the alphabetical form of the binomial and the compositional form, $cos_{\neg\alpha}$ is the cosine similarity between the embeddings of the nonalphabetical form of the binomial and the compositional form, h_{α} and $h_{\neg\alpha}$ are the embeddings of the holistic form of the binomial in alphabetical and nonalphabetical forms respectively (in a sentence context), and c is the embeddings of the compositional form. Since c represents the mean of the embeddings for each word in the binomial out of context, order does not matter. Cosine similarity ranges from -1 to 1 where 1 indicates two extremely similar vectors and -1 indicates two extremely dissimilar vectors.

$$\cos\alpha = \frac{h_{\alpha} \cdot c}{\|h_{\alpha}\| \, \|c\|} \tag{5.1}$$

$$\cos\neg\alpha = \frac{h_{\neg\alpha} \cdot c}{\|h_{\neg\alpha}\| \, \|c\|} \tag{5.2}$$

For each binomial, we then calculated $LogCosSim$ which is the logged quotient of cos_{α} and $cos_{\neg\alpha}$ (Equation 5.3). A larger positive value indicates a greater degree of similarity between the holistic embeddings for the alphabetical form and the embeddings of the compositional form (i.e., the holistic embeddings of the alphabetical form are more similar to the embeddings of the compositional form than the holistic embeddings of the nonalphabetical form are) and a larger negative value represents the opposite.

$$LogCosSim = log(\frac{cos_{\alpha}}{cos_{\neg\alpha}}) \tag{5.3}$$

Analysis

We used a Bayesian mixed-effects model to examine how the semantic similarity between the holistic embeddings and the compositional embeddings trade off as a function of relative and overall frequency. Specifically, we modeled $LogCosSim$ as a function of overall frequency, which was centered and logged, $RelFreq$ which ranged from -0.5 to 0.5 (with 0.5 representing a binomial that appears only in the alphabetical form, and -0.5 representing a binomial that appears only in the nonalphabetical form), and their interaction. Our model is presented below in Equation 5.4.

$$LogCosSim \sim OverallFreq * RelFreq \tag{5.4}$$

5.2.2. Results

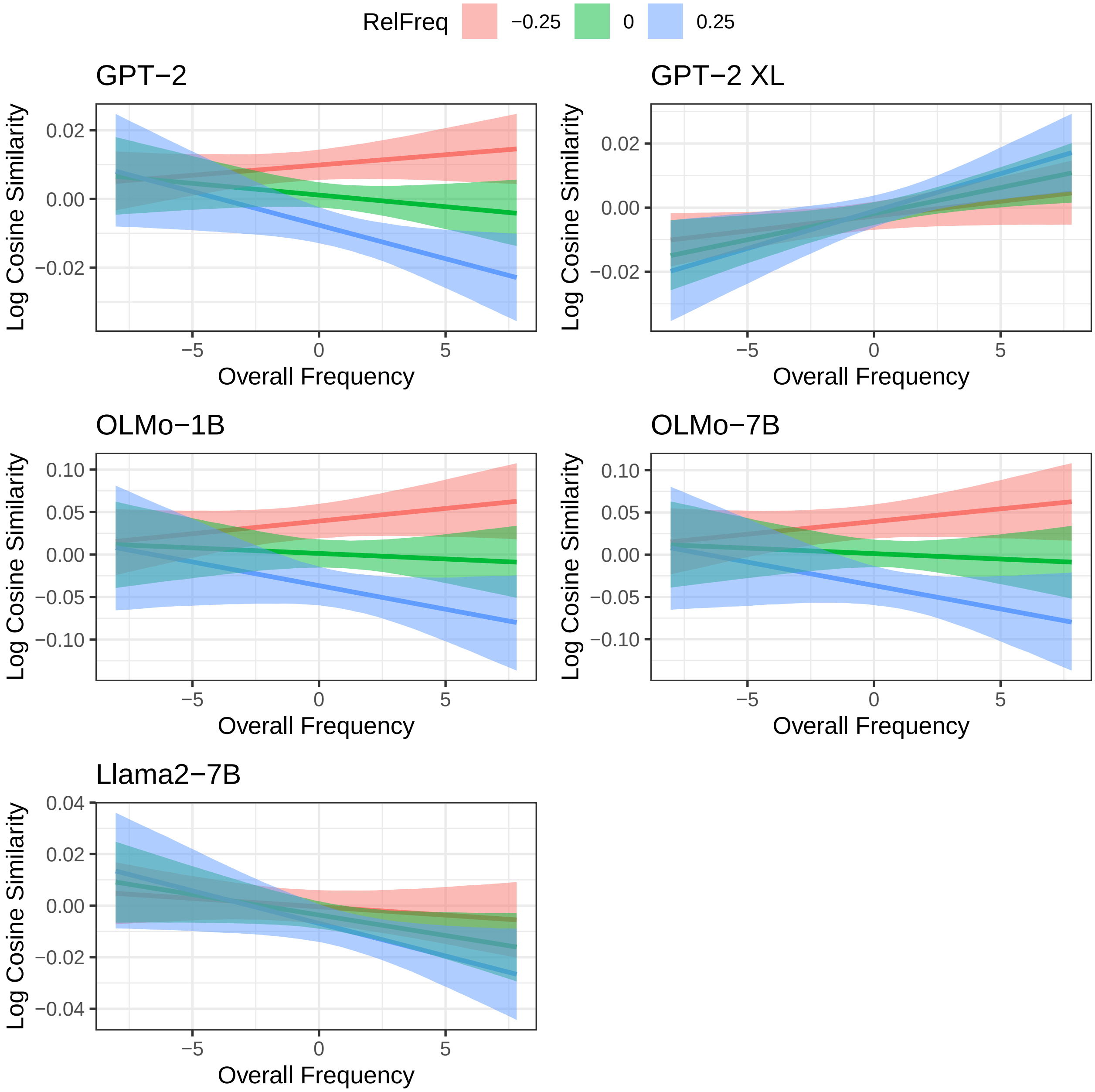
Our results for each model are presented in Table 5.2.1 and visualized in Figure 5.2.1. Following (Houghton et al., 2024) we also report the percentage of posterior samples greater than zero. Since we are using Bayesian mixed-effects models, we are not forced into a binary of significant or non-significant. By reporting the percentage of posterior samples greater than zero, we present a more nuanced picture of our results.

Overall we find a negative effect of relative frequency for GPT-2, OLMo-1B, OLMo-7B, and Llama-2 7B. These results suggest that in general for those models, the embeddings for ordering of the binomial that is more frequent (i.e., higher relative frequency) are less similar to the compositional embeddings than the embeddings for the ordering of the binomial that is less frequent are. Additionally,

Table 5.2.1.: Bayesian linear mixed-effects model results of each model

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
GPT-2					
Intercept	0.00	0.00	0.00	0.00	72.44
OverallFreq	0.00	0.00	0.00	0.00	14.22
RelFreq	-0.04	0.01	-0.05	-0.02	0.00
OverallFreq:RelFreq	0.00	0.00	-0.01	0.00	0.04
GPT-2 XL					
Intercept	0.00	0.00	0.00	0.00	14.77
OverallFreq	0.00	0.00	0.00	0.00	99.67
RelFreq	0.00	0.01	-0.01	0.01	68.45
OverallFreq:RelFreq	0.00	0.00	0.00	0.01	97.82
OLMo-1B					
Intercept	0.00	0.01	-0.01	0.02	56.38
OverallFreq	0.00	0.00	-0.01	0.00	33.00
RelFreq	-0.15	0.03	-0.21	-0.10	0.00
OverallFreq:RelFreq	-0.02	0.01	-0.03	0.00	0.43
OLMo-7B					
Intercept	0.00	0.01	-0.01	0.02	55.77
OverallFreq	0.00	0.00	-0.01	0.00	32.45
RelFreq	-0.15	0.03	-0.21	-0.09	0.00
OverallFreq:RelFreq	-0.02	0.01	-0.03	0.00	0.58
Llama2-7B					
Intercept	0.00	0.00	-0.01	0.00	9.25
OverallFreq	0.00	0.00	0.00	0.00	3.40
RelFreq	-0.01	0.01	-0.03	0.01	8.06
OverallFreq:RelFreq	0.00	0.00	-0.01	0.00	3.17

Figure 5.2.1.: Visualization of the effects of overall frequency and relative frequency on the cosine similarity between embeddings.



for these models there is a negative interaction effect. This suggests that for high-frequency binomials there is a stronger effect of relative frequency than for low-frequency binomials. Specifically, the difference between the embeddings for high relative frequency binomials and the compositional embeddings is even greater in high-frequency binomials than low-frequency ones.

We also find mixed results for overall frequency, with GPT-2 XL showing a strong effect of overall frequency such that the embeddings for high-frequency binomials (ignoring relative frequency) are more similar to the compositional embeddings than the embeddings for low-frequency binomials are.

Finally, the interaction effect between relative frequency and overall frequency are also in the opposite direction for GPT-2 XL compared to the others. Specifically, for GPT-2 XL, as overall frequency increases, the embeddings for the more frequent ordering of the binomials are *more* similar to the compositional embeddings than the embeddings for the less frequent ordering are.

5.2.3. Discussion

Overall we find that for GPT-2, OLMo-1B, OLMo-7B, and Llama2-7B, the representations for more the more frequent form of the binomial diverges more from the representation of the compositional form as a function of overall frequency. That is, for low-frequency binomials, there is not much of a difference between the alphabetical and nonalphabetical orderings of binomials, however for high-frequency binomials, there is a larger difference between the representation of the more frequent ordering and the representation of the compositional ordering.

Interestingly, this is not the case for GPT-2 XL, where the opposite pattern is observed: as overall frequency increases, the representations of the more frequent ordering of the binomial become even more similar to the compositional representations. This suggests that not all large language models are learning the same preferences and further that these preferences may not be completely necessary to generate human-like text.

In summary, our results suggest that for higher frequency binomials in most large language models, the semantic representation for the more frequent form of the binomial diverges more from the representation of the compositional form. This suggests that some large language models tend to learn different representations for high-frequency binomials, similar to what has been argued that humans do (Morgan & Levy, 2016a). However, it's unclear on what timescale this emerges and at what hidden layers this result holds for. For example, does this difference emerge early in training or does it take a large amount of training for these different representations to emerge? Further, since different layers have been proposed to correspond to different functions [e.g., earlier layers may represent more phonological knowledge while later layers may represent more semantic knowledge; Tenney et al. (n.d.)], it is possible that these results may be vary across different layers. In Experiment 2 we examine both of these questions.

5.3. Experiment 2

Experiment 2 is an exploratory analysis examining how representations for binomials emerge throughout training across different hidden layers. Specifically, since OLMo (Groeneveld et al., 2024) released the model's checkpoints at various stages in the training we can examine how our results in Experiment 1 emerge throughout training. Further, since the model is open access we can also examine the different hidden-layers of the model.

5.3.1. Methods

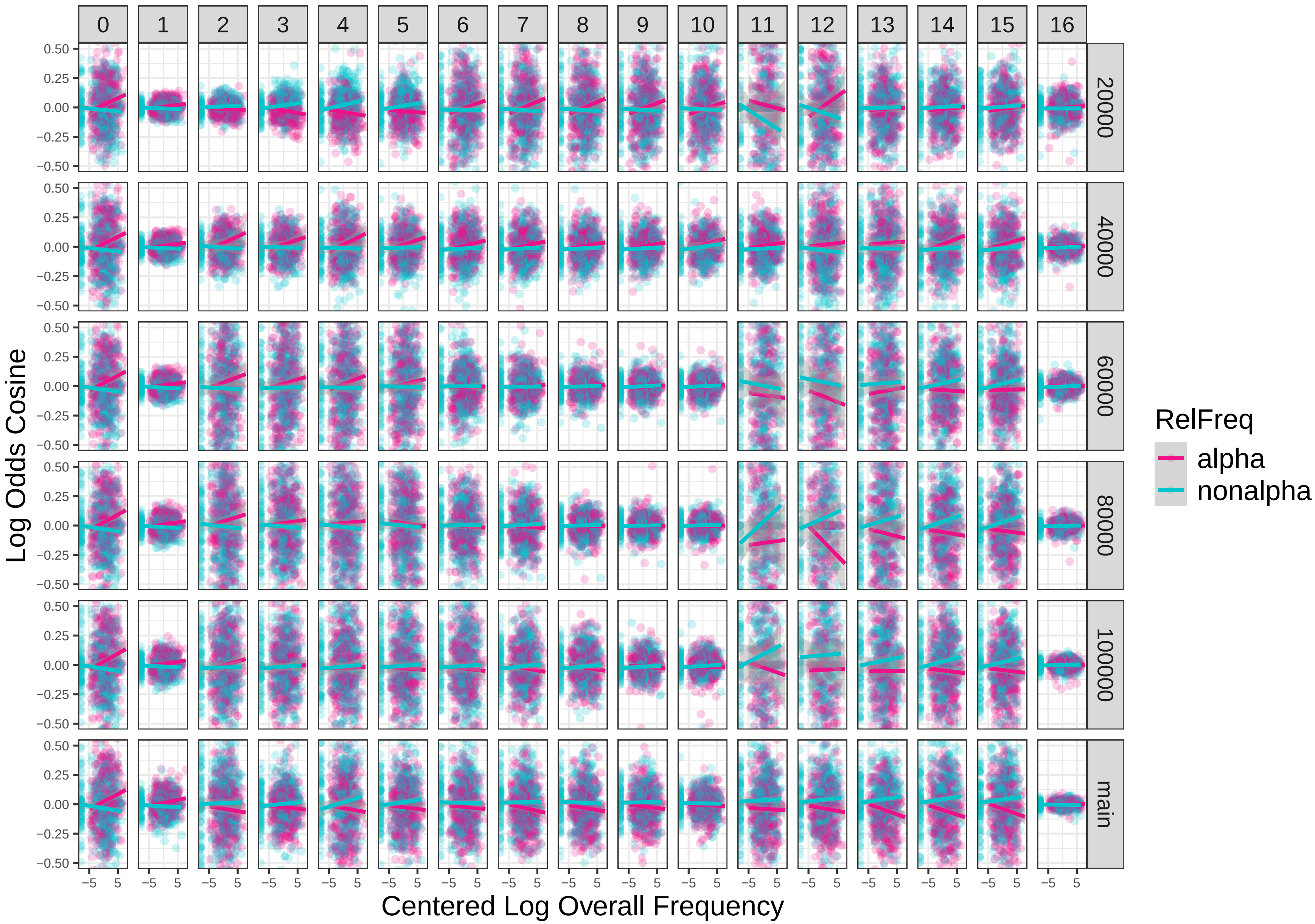
The methods in Experiment 2 were almost identical to those used in Experiment 1, with two main exceptions: first, rather than examining several different large language models, we instead examined a single large language model: OlmO 1B. OlmO 1B has released checkpoints at different stages in learning. As such, we can examine the representations of binomials at different stages of learning. Second, we also examined the representations at each hidden layer in the model in order to examine how the representation changes across layers.

For the present study, we examine the embeddings for our sentences from Experiment 1 at each hidden layer at multiple different steps in the training. In addition to examining the model after being trained, we also examine the embeddings after being trained for 20000 (84B tokens), 40000 (168B tokens), 60000 (252B tokens), 80000 (336B tokens), and 100000 (419B tokens) steps.

5.3.2. Results

A visualization of the embeddings at different layers and different checkpoints is included in Figure 5.3.1.

Figure 5.3.1.: A visualization of the cosine similarity between the holistic embeddings and the compositional embeddings for each hidden layer of each model checkpoint.



There are two notable trends. First, the earlier layers tend to display an opposite pattern from the later layers, with the more frequent embeddings being more similar to the compositional form. Second, the holistic representation of the frequent ordering diverges from the compositional representation at about the 80000th step (336B tokens). We will discuss the implications of both of these in the discussion section.

5.3.3. Discussion

Our results demonstrate that from early on in the training the frequency difference is reflected in the embeddings in the early layers. Interestingly, however, this is not reflected in the representation at later layers. Instead, the differences in representations emerge in later layers over time.

These results are consistent with the general idea that later layers encode more semantic information, since the pattern of the more frequent embeddings diverging is seen in the later layers, while the earlier layers show the opposite pattern.

Additionally, the embeddings seem to converge in the last layer, suggesting that the different representations may not be important for the task of next-word prediction.

Finally, the fact that the pattern emerges rather slowly (taking several hundred billion tokens of training) suggests that the model must experience the binomial quite a lot in order to learn a separate representation for it. This suggests that the model isn't simply learning a separate representation because the binomial occurs in different semantic contexts (because if that were the case we would see the pattern emerge quite early on), but because it occurs frequently. This is in line with usage-based theories that have argued that holistic representations in humans emerge as a function of usage (e.g., Bybee, 2003).

5.4. Conclusion

The present study demonstrates that the semantic embeddings for the more frequent ordering of a given binomial become less similar to the compositional embeddings as a function of the overall frequency of the binomial. That is, the embeddings of the more frequent ordering of a high-frequency binomial (e.g., *bread and butter*) are less similar to the compositional embeddings than the embeddings of the less frequent (e.g., *butter and bread*) are. Another way to frame these results is that the same form (i.e., the same words) can give rise to quantitatively (but systematically) different representations in large language models and this varies depending on the overall frequency of that form (in either order).

It may not seem particularly surprising that the more frequent form diverges in semantic representation from the compositional form. After all, by definition a large language model has more experience with the more frequent form, which means the embeddings are being updated more often for the more frequent ordering. This in turn creates more opportunities for those embeddings to diverge from the compositional embeddings. However, what is interesting is how this effect emerges over time: early on in the training, the embeddings for the more frequent form are more similar to the compositional form across both earlier and later layers. Further, as training continues this stays the case for early layers, but undergoes a reversal in later layers.

One possible explanation for our results is that the more frequent form may be occurring in particularly different contexts from the compositional and less frequent forms (e.g., perhaps they are more idiomatic, such as *black and white*[4]). However, if this were the case then we would expect to see the embeddings for the frequent form to diverge from the embeddings of the compositional form quite early in training. Instead, however, we actually see the opposite early in the training: the embeddings for the more frequent form are *more* similar to those of the compositional form and it takes time for these embeddings to diverge.

[4]Although all of our sentences were sentences that encouraged a compositional reading of our binomials, and very few of our binomials had a particularly idiomatic meaning to begin with.

Another possibility is that early in training for high-frequency binomials, the large language model's experience with the individual words may largely overlap with the large language model's experience with the frequent form of the binomial (e.g., the model's experience with contexts containing the binomial *bread and butter* are also contributing to the large language model's experience with the individual words). Thus, initially these embeddings may be similar until the large language model experiences enough data to learn different representations. As the model experiences more sentence contexts with the binomial, the representation for the more frequent ordering has more opportunities to diverge from the representation of the individual words. This process explains why the same form can give rise to different representations.

Finally, our results can also be considered predictions for how humans may learn representations. Future work would do well to examine whether it is also the case that the semantic representations for the more frequent ordering of high-frequency binomials diverge more from the compositional representations in humans. Our results also make predictions about the timescale of learning: for young children, the pattern of results may actually be the opposite from adults, since at earlier checkpoints in our model the embeddings for the more frequent ordering of high-frequency binomials were more similar to the compositional embeddings.

Chapter 6.

Frequency-dependent preference extremity arises from a noisy-channel processing model

6.1. Introduction

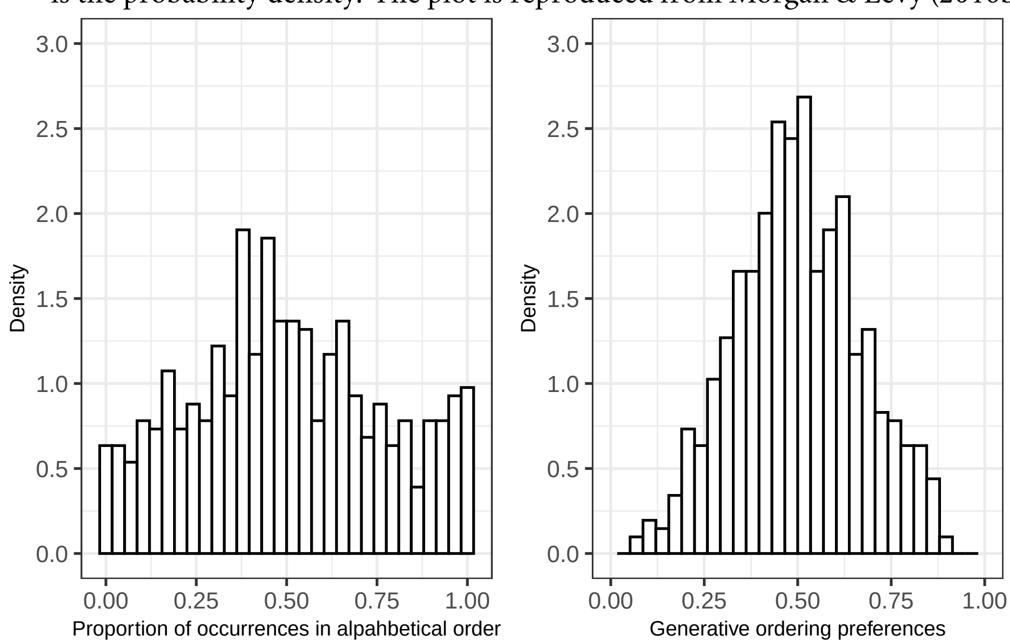
Speakers are often confronted with many different ways to express the same meaning. A customer might ask whether a store sells *radios and televisions* but they could have just as naturally asked whether the store sells *televisions and radios*. However, despite conveying the same meaning, speakers sometimes have strong preferences for one choice over competing choices [e.g., a preference for *men and women* over *women and men*; Benor & Levy (2006); Morgan & Levy (2016a)]. These preferences are driven to some extent by generative preferences (e.g., preference for short words before long words), however they are sometimes violated by idiosyncratic preferences [e.g., *ladies and gentlemen* preferred despite a general men-before-women generative preference; Morgan & Levy (2016a)].

Interestingly, ordering preferences for certain constructions, such as binomial expressions, are often more extreme for higher frequency items (e.g., *bread and butter*). That is, higher-frequency items typically have more polarized preferences (Liu & Morgan, 2020, 2021; Morgan & Levy, 2015, 2016a, 2016b). This phenomenon is called *frequency-dependent preference extremity*, and while there is evidence of it in several different constructions, it is still unclear what processes this phenomenon is driven by. For example, it could be a consequence of learning processes or a consequence of sentence processing more broadly. In the present paper we examine whether a noisy-channel processing model (Gibson, Bergen, et al., 2013) combined with transmission across generations (Kirby et al., 2008; Reali & Griffiths, 2009) can account for frequency-dependent preference extremity.

6.1.1. Frequency-Dependent Preference Extremity

Frequency-dependent preference extremity has been documented for a variety of different constructions in English (Liu & Morgan, 2020, 2021; Morgan & Levy, 2015, 2016b). For example, Morgan & Levy (2015) demonstrated that more frequent binomial expressions (e.g., *bread and butter*) are more polarized (i.e., are preferred in one order overwhelmingly more than the alternative). These ordering preferences are also not simply a result of generative ordering preferences [e.g., short words before long words; Morgan & Levy (2016a)]. Interestingly, Morgan & Levy (2016b) even showed that the distribution of binomial orderings at the corpus-wide level are different than what would be expected given the generative preferences for the binomials (see Figure 6.1.1).

Figure 6.1.1.: The left plot is a plot of the relative orderings of binomials in the corpus data from Morgan & Levy (2015), the right is the plot of the generative preferences of binomials in the same corpus. The x-axis is proportion of occurrences in alphabetical order and the y-axis is the probability density. The plot is reproduced from Morgan & Levy (2016b).



Additionally, Liu & Morgan (2020) demonstrated that the dative alternation in English also shows evidence of frequency-dependent preference extremity (e.g., *give the ball to him* vs *give him the ball*). Specifically, they demonstrated that higher frequency verbs have more polarized preferences with respect to the dative alternation. Similarly, Liu & Morgan (2021) showed that in adjective-adjective-noun constructions, the adjective orderings also show frequency-dependent preference extremity. That is, adjectives in adjective-adjective-noun constructions with higher overall frequencies (i.e., the summed counts of both orderings) show stronger ordering preferences, even after taking into account generative preferences of adjective orderings.

Interestingly, frequency-dependent preference extremity patterns differently from rule-following regularization processes (e.g., morphological regularization) where it is the low-frequency items that become more regular [rather than the high-frequency items; Singleton & Newport (2004)]. For example, Schneider et al. (2020) demonstrated through a noisy-channel processing model that regularization can arise from learners attributing variation in the low-frequency items to noise. On the other hand, frequency-dependent preference extremity patterns more similarly to other processes, such as semantic entrenchment, where it is the high-frequency items that develop strict preferences (Harmon & Kapatsinski, 2017; Theakston, 2004). For example, people are generally more willing to accept a low-frequency intransitive verb in a transitive context than a high-frequency intransitive verb [e.g., *He vanished it* is judged to be more acceptable than *He disappeared it*; Kapatsinski (2018); Robenalt & Goldberg (2015); Theakston (2004)].

Why is it that it is the high-frequency items that develop more polarized preferences in frequency-dependent preference extremity? One possibility is that it occurs as an interaction between imperfect learning and transmission across generations. For example, it's possible that while learners of a language are in general very good at learning the statistical patterns in the language (e.g., Saffran et al., 1996; Yu & Smith, 2007), they may do so imperfectly and with a bias towards preference extremity. If a learner hears 70 tokens of *bread and butter* and 30 tokens of *butter and bread*, they may imperfectly infer the ordering preference and transmit the language with a more skewed distribution (e.g., 75 tokens *bread and butter* and 25 tokens of *butter and bread*). Indeed, previous studies have shown that learners will reproduce the more frequent item at an even higher rate than they heard it (Harmon & Kapatsinski, 2017; Hudson Kam & Newport, 2009). As the language is transmitted from generation to generation, it is possible this compounds until the highest-frequency items develop polarized ordering preferences.

Following this logic, Morgan & Levy (2016b) investigated whether frequency-dependent preference extremity can arise as a result of imperfect learning across generations. They found that a data generation model with a frequency-*independent* bias can result in frequency-*dependent* preference extremity across generations of learners in a 2-alternative iterated learning paradigm. They argued that frequency-dependent preference extremity emerges because for low-frequency items, the preference extremity bias cannot overcome the learner's generative preferences for maintaining variation, but for high-frequency items, it can. In other words, for lower frequency items, learners may rely more on their generative preferences because they haven't heard the item very much. As the language is transmitted across many generations, it may result in frequency-dependent preference extremity.

While there is good evidence that a frequency-*independent* preference extremity bias can account for frequency-dependent preference extremity across generations, it remains unclear what processes in language transmission are analogous to this preference extremity bias.

6.1.2. Noisy-Channel Processing

One possibility is that the frequency-independent preference extremity bias is a product of noisy-channel processing (Gibson, Bergen, et al., 2013). Listeners are confronted with a great deal of noise in the form of perception errors (e.g., a noisy environment) and even production errors [speakers don't always say what they intended to; Gibson, Bergen, et al. (2013)]. In order to overcome these errors, a processing system must take into account the noise of the system, for example by probabilistically determining whether the perceived utterance was infact intended by the speaker.

Indeed, there is evidence that our processing system does take noise into account. For example, Ganong (1980) found that people will process a non-word as being a word under noisy conditions. Additionally, Felty et al. (n.d.) demonstrated that when listeners do misperceive a word, the word that they believe to have heard tends to be higher frequency than the target word. Further, Keshev & Meltzer-Asscher (2021) found that in Arabic, readers will even process ambiguous subject/object relative clauses as the more frequent interpretation, even if this interpretation compromises subject-verb agreement. These results taken together suggest that misperceptions may sometimes actually be a consequence of noisy-channel processing (although it's worth noting that good-enough processing theories also make very similar predictions, e.g., Ferreira & Patson, 2007).

Further, people will even process *grammatical* utterances, as a more frequent or plausible interpretation (Christianson et al., 2001; Levy, 2008; Poppels & Levy, 2016). This can even arise in two interpretations that cannot both be consistent with the original sentence. For example, Christianson et al. (2001) demonstrated that when people read the sentence *While the man hunted the deer ran into the woods*, people will answer in the affirmative for both *Did the man hunt the deer?* and *Did the dear run into the woods?*. Levy (2008) argued that this phenonenon was explained by noisy-channel processing, since a single insertion results in plausible, grammatical constructions for both meanings (*While the man hunted it the deer ran into the woods* vs *While the man hunted the deer it ran into the woods*).

In order to account for findings like these, Gibson, Piantadosi, et al. (2013) developed a computational model that demonstrated how a system might take into account noise (see Levy, 2008 for a similar approach). Specifically, their model operationalizes noisy-channel processing as a Bayesian process where a listener estimates the probability of the speaker's intended utterance given what they perceived. Specifically, this is operationlized as being proportional to the prior probability of the intended utterance multiplied by the probability of the intended utterance being corrupted to the perceived utterance (See Equation 6.1):

$$P(S_i|S_p) \propto P(S_i)P(S_i \rightarrow S_p) \tag{6.1}$$

where $P(S_i|S_p)$ is the probability of the intended utterance given the perceived utterance, $P(S_i)$ is the prior probability of the intended utterance, and $P(S_i \rightarrow S_p)$ is the probability of the perceived utterance (S_p) given the intended utterance (S_i). If the perceived utterance is *butter and bread*, for example, the listener can infer the probability that the intended utterance was *bread and butter* or *butter and bread*.

Gibson, Piantadosi, et al. (2013)'s model made a variety of interesting predictions. For example, the model predicted that when people are presented with an implausible sentence (e.g., *the mother gave the candle the daughter*), they should be more likely to interpret the plausible version of the sentence (e.g., *the mother gave the candle to the daughter*) if there is increased noise (e.g., by adding syntactic errors to the filler items, such as a deleted function word). Their model also predicted that increasing the likelihood of implausible events (e.g., by adding more filler items that were implausible, such as *the girl was kicked by the ball*) should increase the rate of implausible interpretations of the sentence. Interestingly both of these results were born out in their experimental data. In a follow up study, Poppels & Levy (2016) further demonstrated that word-exchanges (e.g., *The ball kicked the girl* vs *The girl kicked the ball*) are also taken into account by comprehenders. These results taken together suggest that humans do utilize a noisy-channel system in processing.

In addition to Gibson, Piantadosi, et al. (2013), previous research has demonstrated that noisy-channel processing models may also account for certain types of regularization (e.g., Ferdinand et al., 2019; Schneider et al., 2020). For example, as mentioned earlier, Schneider et al. (2020) demonstrated that a noisy-channel model can account for some rule-following regularization processes (e.g., morphological regularization). However, it is unclear whether noisy-channel processing models can also account for frequency-dependent preference extremity.

6.1.3. Present Study

Given the evidence of noisy-channel processing, it is possible that the frequency-dependent preference extremity that Morgan & Levy (2016b) saw is a product of listeners' noisy-channel processing. Perhaps when learners hear the phrase *butter and bread*, they think the speaker intended *bread and butter*, which results in an activation of *bread and butter* even though they didn't hear it. This activation could potentially even be stronger for *bread and butter* than *butter and bread* in cases where the listener thinks the speaker made a mistake. Further, this may compound over time for high frequency items, but not for low frequency items. Thus, the present study examines whether Gibson, Piantadosi, et al. (2013)'s noisy-channel processing model can also predict frequency-dependent preference extremity across generations of language transmission.

6.2. Dataset

Following Morgan & Levy (2016a), we use Morgan & Levy (2015)'s corpus of 594 Noun-Noun binomial expressions (e.g., *bread and butter*). There is evidence that human binomial ordering preferences are driven by a combination of generative preferences and observed preferences. Generative preferences are abstract constraints on ordering preferences, such as a preference for short words before long words, or male-coded terms before female-coded terms. The observed preference for a given binomial is the percentage that a given binomial occurs in alphabetical vs nonalphabetical form. That is, if *cats and dogs* appears 40 times in a corpus, and *dogs and cats* appears 60 times, then the observed preference for the alphabetical form is 0.4. The corpus also contains the overall frequency (total count of alphabetical and nonalphabetical forms for a given binomial) which has been shown to affect the strength of ordering preferences (Morgan & Levy, 2016a). A detailed description of the constraints is listed below:

1. The estimated generative preferences for each binomial, which are values between 0 and 1 representing the alphabetical ordering preferences (a neutral reference order), estimated from various phonological and semantic features that are known to influence binomial ordering preferences (Morgan & Levy, 2015). The generative constraints are calculated using Morgan & Levy (2015)'s model. Values closer to zero represent a generative preference for the nonalphabetical order, while values closer to 1 represent a generative preference for the alphabetical order.

2. The observed binomial orderings preferences (hereafter: observed preferences) which are the proportion of binomial orderings that are in alphabetical order for a given binomial. A visualization of the distribution of observed preferences and generative preferences is included below in Figure 6.1.1.

3. The overall frequency of a binomial expression (the frequency of AandB plus the frequency of BandA). Frequencies were obtained from the Google Books *n*-grams corpus (Lin et al., 2012), which is orders of magnitude larger than the language experience of an individual speaker, and thus provides reliable frequency estimates for these expressions.

6.3. Model

Following Morgan & Levy (2016b), we use a 2-alternative iterated learning paradigm. In our iterated learning paradigm, at each generation, learners hear N tokens of a given binomial with some in alphabetical (AandB) and some in nonalphabetical (BandA) order. The learner's goal is to learn the ordering preferences for each binomial. After hearing all N tokens, the learner then produces N tokens to the next generation. This process then repeats. Morgan & Levy (2016b) used a beta-binomial model: A learner has some prior over binomial ordering preferences, which can be expressed as pseudocounts favoring each order (e.g.~3 pseudocounts for AandB and 7 for BandA). Each time the learner hears a binomial, they update their beliefs by adding 1 count to the perceived order, e.g., if they heard AandB, adding 1 AandB count. We modify this by instead having the learner update their beliefs in proportion to what they believe the intended order was: e.g., if they believe the intended utterance was AandB with 50% probability and BandA with 50% probability, they will add 0.5 to each count. These updated beliefs then influence their beliefs about future intended utterances (Equation 6.1).

Specifically, the prior probability over the binomial ordering preferences, ($P(S_i)$), follows Equation 6.2 and Equation 6.3. α_1 and α_2 are pseudocounts of the alphabetical and nonalphabetical forms respectively.

$$S_i \sim Bernoulli(p_{theta}) \tag{6.2}$$

$$p_{theta} \sim Beta(\alpha_1, \alpha_2) \tag{6.3}$$

After hearing a token, learners compute $P(S_i = AandB|S_p)$ according to Equation 6.1. $P(S_i \rightarrow S_p)$ is determined by a fixed noise parameter, which we will call p_{noise}. p_{noise} represents the learner's belief of how likely a binomial ordering is to have been swapped (i.e., AandB being swapped to BandA or vice versa).

To initialize p_{theta}, and thus $P(S_i)$, before the learner hears any data, we used the mean and concentration parametrization of the beta distribution. The mean (μ) represents the expectation of the distribution (the mean value of draws from the distribution). The concentration parameter (ν) describes how dense the distribution is. Before the learner hears any data, μ is equal to the generative preference for the binomial (taken from Morgan & Levy, 2016b). ν is a free parameter, set to 10 for all simulations in this paper.[1] α_1 and α_2 can also be expressed in terms of μ and ν:

$$\alpha_1 = \mu \cdot \nu \tag{6.4}$$

$$\alpha_2 = (1 - \mu) \cdot \nu \tag{6.5}$$

For all future tokens, learners will use the updated $P(S_i)$ from the previous token, where $P(S_i = AandB)$ is the expectation of p_θ. Crucially, this value will be different for each token of learning due to the update that occurs on the previous token.

[1] Changing ν does not qualitatively change the pattern of the results for any simulations in the paper, as long as it's greater than 2.

$$P(S_i = AandB) = \mathbb{E}(p_\theta) \tag{6.6}$$

We then use $P(S_i)$ and p_{noise} to compute $P(S_i|S_p)$, following Equation 6.1. If the perceived binomial is alphabetical (AandB), we compute the unnormalized probability of the alphabetical and nonalphabetical orderings according to the below equations. Note that the process is comparable if the perceived binomial is nonalphabetical.

$$P_{raw}(S_i = AandB|S_p = AandB) = P(S_i = AandB) \cdot (1 - p_{noise}) \tag{6.7}$$

$$P_{raw}(S_i = BandA|S_p = AandB) = (1 - P(S_i = AandB)) \cdot p_{noise} \tag{6.8}$$

After calculating the unnormalized (raw) probabilities, they are then normalized:

$$\hat{p}_\alpha = \frac{P_{raw}(S_i = AandB|S_p = AandB)}{P_{raw}(S_i = AandB|S_p = AandB) + P_{raw}(S_i = BandA|S_p = AandB)} \tag{6.9}$$

$$\hat{p}_{\neg\alpha} = 1 - \hat{p}_\alpha \tag{6.10}$$

where $\hat{p}\alpha$ is the probability that the intended binomial order was the alphabetical order, and $\hat{p}\neg\alpha$ is the probability that the intended binomial order was the nonalphabetical order.

We then update α_1' and $\alpha 2'$to be used as the parameters of $p\theta$, and thus $P(S_i)$, when the learner hears the next token. This update is done according to the following equation:

$$\alpha_1' = \alpha_1 + \hat{p}_{\alpha} \tag{6.11}$$

$$\alpha_2' = \alpha_2 + \hat{p}_{\neg\alpha} \tag{6.12}$$

Note that when the learner hears any binomial, they update their beliefs about the probability of both the alphabetical *and* nonalphabetical forms of the binomial (in proportion to how likely they believe each ordering was intended by the speaker).

When the learner is done hearing N tokens and updating their beliefs of $P(S_i)$ for a given binomial, they then produce N tokens for the next generation of learners. These are generated binomially, where $\theta = \mathbb{E}(p_{\theta})$ is the inferred probability of the alphabetical form of a given binomial. For the first generation of speakers (before any learning has occurred), θ is initialized at 0.5.

When producing each token, there is also a possibility that the speaker makes an error and produces an unintended ordering of the binomial. The speaker error is analogous to a speaker choosing to produce a binomial ordering (AandB or BandA), and then accidentally flipping it. For example, perhaps they intended to say *butter and bread*, but accidentally said *bread and butter* (or vice versa). Note that the "unintended ordering" is whichever order the speaker did not choose to produce on that trial, regardless of the overall preference for the binomial. In order to model this, the speaker produces a token in the unintended order with probability $p_{SpeakerNoise}$. This is a fixed parameter in the model and remains constant across binomials and generations.

This process continues iteratively for $ngen$ generations.

6.4. Results

We present our results in two main sections. The first section demonstrates the effects of the speaker and listener noise parameters (p_{noise} and $p_{SpeakerNoise}$ respectively) on simulations of individual binomials. The aim of this section is to examine whether the model can account for frequency-dependent preference extremity across individual binomials varying in frequency.

The second section compares our model's predicted binomial orderings across a range of binomials to the real-world corpus-wide distribution. In this section, rather than simulating individual binomials, we simulate the distribution of binomial orderings across the entire dataset of binomials from Morgan & Levy (2015) with the intent of examining whether our model can capture the corpus-wide distribution.

6.4.1. Speaker vs Listener Noise

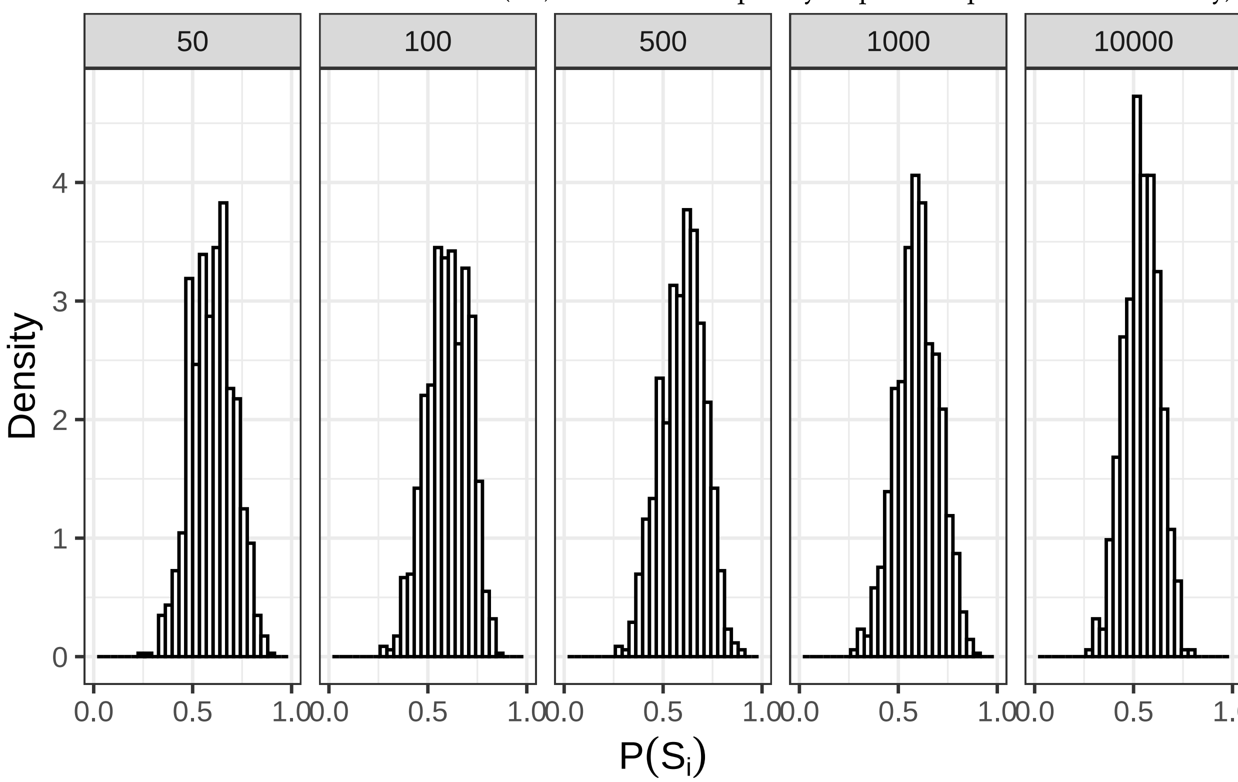
First we demonstrate that frequency-dependent preference extremity does not arise when there is no listener or speaker noise.[2] Instead we see convergence to the prior, which is expected following Griffiths & Kalish (2007). They demonstrated that when learners sample from the posterior in an iterated learning paradigm, the stationary distribution converges to the prior. To confirm this, we simulated the evolution of a single binomial across 500 generations with various N (50, 100, 500, 1000, and 10,000). The generative preference was 0.6. 1000 chains were run. We then examined the model's inferred ordering preference in the final generation. A visualization of the results is presented in Figure 6.4.1.

We then systematically manipulated N, listener noise (p_{noise}) and speaker noise ($p_{SpeakerNoise}$). Specifically, we varied N across 100, 1000, and 10000, and listener and speaker noise were varied across 0, 0.033, 0.066, and 0.1. We ran simulations for every combination of these values (Figure 6.4.2). For these simulations, the generative preference was set to 0.6 and 1000 chains were run across 500 generations.

Our results suggest that frequency-dependent preference extremity does arise from the model when noise is introduced, but only if listener noise is greater than speaker noise. Further

[2] All code and results can be found publicly available here: https://github.com/znhoughton/Noisy-Channel-Iterated-Learning

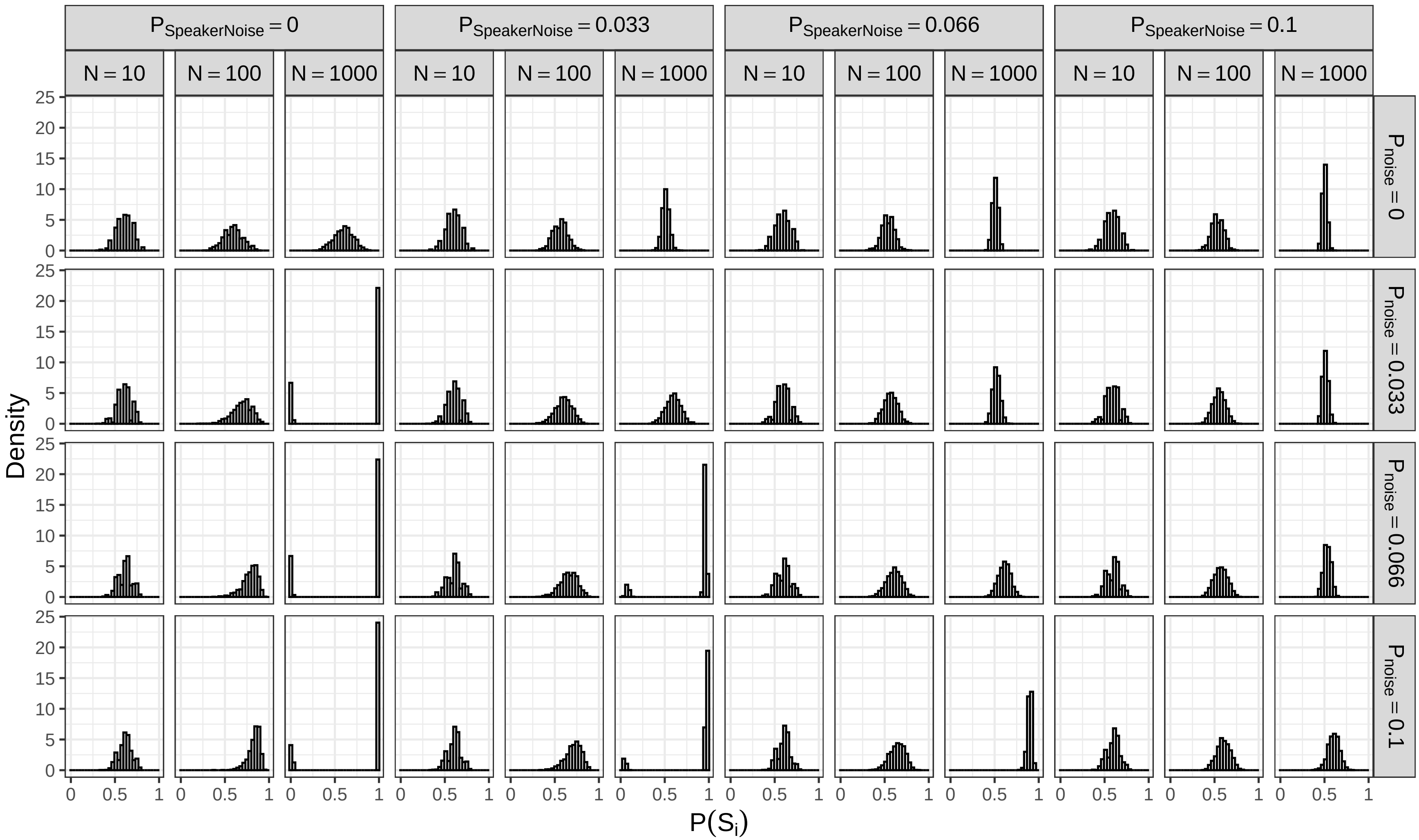
Figure 6.4.1.: A plot of the distribution of simulated binomials at the 500th generation, varying in frequency. The top value represents N, which is the overall frequency of a binomial regardless of ordering (i.e., count(AandB) + count(BandA)). On the x-axis is the predicted probability of producing the binomial in alphabetical form. On the y-axis is probability density. Speaker and listener noise was set to 0. The generative preference was 0.6, and nu was set to 10. 1000 chains were run. Note that all values of N produce dense distributions clustered around 0.6 (i.e., there is no frequency-dependent preference extremity).



our results demonstrate that if listener noise is greater than speaker noise, then the greater the difference between the listener and speaker noise, the stronger the preference extremity effect (this is demonstrated by moving vertically down the column labeled $p_{SpeakerNoise} = 0$ in Figure 6.4.2).

Interestingly this preference extremity disappears if the listener's noise parameter is less than or equal to the speaker's noise parameter. For example, notice how if you split the plot along the diagonal, all the plots on the top half, including the diagonal, show no evidence of preference extremity. These graphs are all visualizations where the speaker noise is greater than or equal to the listener noise.

Figure 6.4.2.: Our simulation results for every combination of speaker noise, listener noise, and N. Note that there is an increase in ordering preference extrmity as N increases when listener noise is greater than speaker noise. N corresponds to the overall frequency of the binomial (count of AandB plus count of BandA) and varies across 10, 100, and 1000. Both speaker and listener noise were varied across 0, 0.033, 0.066, and 0.1. The distributions in the plot demonstrate the inferred ordering preference at the 500th generation.



It is useful to revisit here what the speaker and listener noise parameters represent. The speaker noise parameter is how often the speaker produces an error and the listener noise parameter is the listeners' belief of how noisy the environment is. Note that a speaker error here is not whether the speaker produces the more frequent binomial ordering, but rather whether the speaker produces the intended binomial ordering. In other words, if a speaker intends to produce *butter and bread*, and instead produces *bread and butter*, this is an error in our model. Framed this way, one explanation for our results is that when the listener is inferring more noise than the speakers are producing, they are relying more on their inferences, which can become more and more extreme. On the other hand, if they're not inferring enough noise, then they are relying more on the data. The greater the speaker noise, due to how we operationalized speaker noise, the more balanced the data will be.

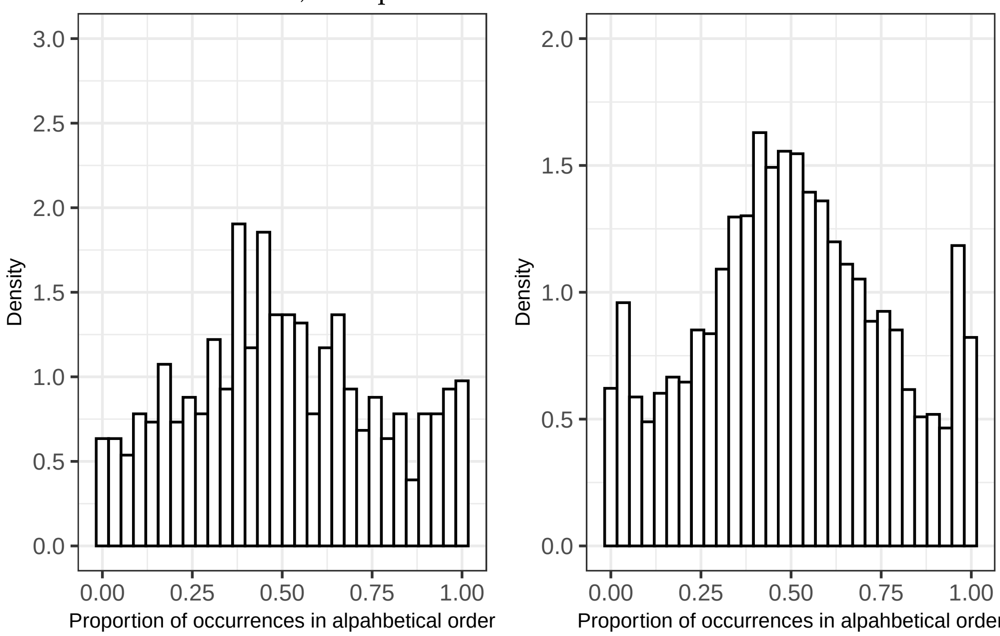
Thus our model makes a novel prediction: In order to account for frequency-dependent preference extremity, listeners must be inferring more noise than speakers are actually producing.

6.4.2. Corpus Data

Finally, we now demonstrate that our model also predicts the language-wide distribution of binomial preference strengths seen in the corpus data. In order to demonstrate this, we simulated model predictions for all 594 binomials from Morgan & Levy (2015). The model estimated the ordering preference across 500 generations with 10 chains each. Values for the generative preference and N for each binomial were taken from Morgan & Levy (2015)'s corpus. Listener noise was set to 0.02 and speaker noise to 0.005. Note that we scale N based on an estimated lifetime exposure of 300 million tokens (Levy et al., 2012).

Our results demonstrate that our model can approximate the distribution in the corpus data (See Figure 6.4.3). In other words, the corpus-wide distribution of binomial orderings according to our model is similar to the ordering we see in actual corpus data. Further, the distribution is qualitatively similar regardless of listener and speaker noise parameters, as long as listener noise is greater than speaker noise. Altogether, this suggests that our model both captures the phenomenon of frequency-dependent preference extremity, but also in capturing it our model also predicts a similar distribution of binomial orderings to what we see in corpus data.

Figure 6.4.3.: A plot of the stationary distribution of ordering preferences in the corpus data from Morgan & Levy (2015) and the distribution of ordering preferences after 500 generations of our iterated learning model (left and right respectively). For our simulations, the binomial frequencies and generative preferences were matched with the corpus data. Listener noise was set to 0.02, and speaker noise was set to 0.005.



6.5. Conclusion

The present study examined whether a noisy-channel processing model (Gibson, Piantadosi, et al., 2013) integrated in an iterated learning model (Morgan & Levy, 2016b) can capture the effects of frequency-dependent preference extremity. Our results demonstrate that frequency-dependent preference extremity can emerge from a noisy-channel processing model when listeners infer more noise in the environment than the speakers actually produce. Our results also make novel predictions. For example, if our current model is accurate, it suggests that listeners assume more noise than the speakers produce. Further, it suggests that for high-frequency binomials, such as *butter and bread*, hearing *butter and bread* may activate *bread and butter* more strongly than *butter and bread*. Finally, it seems more unlikely that a speaker would unintentionally produce the unintended ordering for high-frequency binomials than low-frequency binomials (e.g., producing *butter and bread,* when they mean to say *bread and butter)*. Thus it will also be interesting to examine models that don't use a fixed speaker-noise parameter.

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

Bansal, S., Ford, J. M., & Spering, M. (2018). The function and failure of sensory predictions. *Annals of the New York Academy of Sciences*, *1426*(1), 199–220. https://doi.org/10.1111/nyas.13686

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

Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On meaning, form, and understanding in the age of data*. 51855198. https://aclanthology.org/2020.acl-main.463/

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

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407. https://doi.org/10.1006/cogp.2001.0752

Clark, A. (2013). Are we predictive engines?: Perils, prospects, and the puzzle of the porous perceiver. *Behavioral and Brain Sciences*, *36*(3), 233253. https://www.research.ed.ac.uk/en/publications/are-we-predictive-engines-perils-prospects-and-the-puzzle-of-the-

Cooper, W. E., & Ross, J. R. (1975). World order. *Papers from the Parasession on Functionalism*, *11*, 63111.

Felty, R. A., Buchwald, A., Gruenenfelder, T. M., & Pisoni, D. B. (n.d.). Misperceptions of spoken words: Data from a random sample of american english words. *The Journal of the Acoustical Society of America*, *134*, 572–585. https://www.robfelty.com/academic-files/docs/FeltyEtAl2013.pdf

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68. https://doi.org/10.1016/j.cognition.2018.12.002

Ferreira, F., & Chantavarin, S. (2018). Integration and Prediction in Language Processing: A Synthesis of Old and New. *Current Directions in Psychological Science*, *27*(6), 443–448. https://doi.org/10.1177/0963721418794491

Ferreira, F., & Patson, N. D. (2007). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110. https://psycnet.apa.org/record/1981-07020-001

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*(7), 1079–1088. https://doi.org/10.1177/0956797612463705

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

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480. https://doi.org/10.1080/15326900701326576

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. (2018). Colorless green recurrent networks dream hierarchically. *arXiv Preprint 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

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66. https://doi.org/10.1016/j.cogpsych.2009.01.001

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. (2023). Defragmenting learning. *Cognitive Science*, *47*(6), e13301. https://doi.org/10.1111/cogs.13301

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

Keshev, M., & Meltzer-Asscher, A. (2021). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. *Cognitive Psychology*, *124*, 101359. https://doi.org/10.1016/j.cogpsych.2020.101359

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. https://doi.org/10.1073/pnas.0707835105

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

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*.

LeBrun, B., Sordoni, A., & O'Donnell, T. J. (2022). Evaluating distributional distortion in neural language modeling. *arXiv Preprint arXiv:2203.12788*.

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

Levy, R. (2008). *A noisy-channel model of human sentence comprehension under uncertain input*. 234243. https://aclanthology.org/D08-1025.pdf

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., Grant, E., Smolensky, P., Griffiths, T. L., & Linzen, T. (n.d.). *Universal linguistic inductive biases via meta-learning*. https://doi.org/10.48550/arXiv.2006.16324

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. (2015). *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

Pan, D., & Bergen, B. K. (2025). *Are explicit belief representations necessary? A comparison between large language models and bayesian probabilistic models*. https://pages.ucsd.edu/~bkbergen/papers/panbergen2025.pdf

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.

Piantadosi, S. T. (2023). Modern language models refute chomsky's approach to language. *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, 353–414.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv Preprint arXiv:2208.02957*.

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

Poppels, T., & Levy, R. (2016). *Structure-sensitive noise inference: Comprehenders expect exchange errors.* https://tpoppels.github.io/files/2016-poppels-levy-cogsci-proceedings.pdf

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

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

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. https://doi.org/10.1016/j.cognition.2009.02.012

RESCORLA, R. A. (1968). Probability of shock in the presence and absence of cs in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*(1), 1–5. https://doi.org/10.1037/h0025984

Robenalt, C., & Goldberg, A. E. (2015). Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics*, *26*(3), 467–503. https://doi.org/10.1515/cog-2015-0004

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

Schneider, J., Perkins, L., & Feldman, N. H. (2020). *A noisy channel model for systematizing unpredictable input variation*. 533547. http://www.lingref.com/bucld/44/BUCLD44-43.pdf

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

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of american sign language from inconsistent input. *Cognitive Psychology*, *49*(4), 370407. https://www.sciencedirect.com/science/article/pii/S0010028504000295

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.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., et al. (2024). Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv Preprint arXiv:2402.00159*.

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.

Tartaglini, A. R., Feucht, S., Lepori, M. A., Vong, W. K., Lovering, C., Lake, B. M., & Pavlick, E. (2023). Deep neural networks can learn generalizable same-different visual relations. *arXiv Preprint arXiv:2310.09612*.

Tenney, I., Das, D., & Pavlick, E. (n.d.). *BERT rediscovers the classical NLP pipeline*. https://doi.org/10.48550/arXiv.1905.05950

Theakston, A. L. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development*, *19*(1), 15–34. https://doi.org/10.1016/j.cogdev.2003.08.001

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.

Yu, C., & Smith, L. B. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science*, *18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

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

Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., & Zhou, D. (n.d.). *Take a step back: Evoking reasoning via abstraction in large language models*. https://doi.org/10.48550/arXiv.2310.06117

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

Appendix A.

Full Model Results

Table A.0.1.: Model results for each language model. The Estimate is given in the "Est." column, the standard deviation of the posterior is given in the "Err." column. The columns labeled 2.5 and 97.5 represent the lower and upper confidence interval boundaries. AbsPref is the abstract ordering preferences, Observed is the observed preference in corpus data, and Freq is the overall frequency of the binomial.

GPT-2					**GPT-2XL**			
	Est.	**Err.**	**2.5**	**97.5**	**Est.**	**Err.**	**2.5**	**97.5**
Intercept	-0.10	0.10	-0.30	0.10	0.05	0.09	-0.13	0.23
AbsPref	-0.52	0.64	-1.81	0.69	-0.89	0.63	-2.17	0.29
Observed	**4.62**	0.50	3.66	5.59	**5.34**	0.46	4.45	6.25
Freq	-0.04	0.06	-0.15	0.07	-0.01	0.05	-0.11	0.09
AbsPref:Freq	0.10	0.39	-0.66	0.86	-0.17	0.36	-0.87	0.53
Observed:Freq	**0.96**	0.24	0.49	1.43	**1.01**	0.21	0.59	1.43
Llama-2 7B					**Llama-2 13B**			
	Est.	**Err.**	**2.5**	**97.5**	**Est.**	**Err.**	**2.5**	**97.5**
Intercept	0.22	0.13	-0.03	0.47	0.12	0.08	-0.04	0.27
AbsPref	1.11	0.84	-0.40	2.91	0.32	0.54	-0.72	1.38
Observed	**3.07**	0.64	1.81	4.31	**5.25**	0.40	4.46	6.05
Freq	0.04	0.07	-0.10	0.17	-0.08	0.04	-0.16	0.01
AbsPref:Freq	-0.32	0.47	-1.24	0.59	-0.02	0.32	-0.64	0.60
Observed:Freq	0.23	0.28	-0.33	0.78	**0.72**	0.19	0.34	1.09
Llama-3 8B					**Llama-3 70B**			
	Est.	**Err.**	**2.5**	**97.5**	**Est.**	**Err.**	**2.5**	**97.5**
Intercept	0.15	0.09	-0.03	0.33	0.04	0.05	-0.06	0.14
AbsPref	0.23	0.59	-0.92	1.42	0.10	0.38	-0.63	0.85
Observed	**5.64**	0.46	4.75	6.54	**5.00**	0.27	4.49	5.52
Freq	-0.07	0.05	-0.17	0.03	-0.05	0.03	-0.11	0.00
AbsPref:Freq	0.07	0.36	-0.63	0.78	-0.11	0.21	-0.52	0.30
Observed:Freq	**0.60**	0.22	0.18	1.03	**0.65**	0.12	0.41	0.89
OLMo 1B					**OLMo 7B**			
	Est.	**Err.**	**2.5**	**97.5**	**Est.**	**Err.**	**2.5**	**97.5**
Intercept	0.06	0.08	-0.09	0.22	0.04	0.07	-0.10	0.18
AbsPref	0.69	0.54	-0.33	1.79	-0.86	0.51	-1.88	0.11
Observed	**4.36**	0.39	3.58	5.12	**5.37**	0.36	4.67	6.08
Freq	0.06	0.04	-0.02	0.14	0.01	0.04	-0.07	0.08
AbsPref:Freq	-0.12	0.31	-0.73	0.47	0.10	0.28	-0.47	0.64
Observed:Freq	**0.81**	0.19	0.44	1.17	**0.70**	0.17	0.37	1.04

Appendix B.

Individual Constraints at Each Checkpoint

Table B.0.1.: Model results examining the effect of each individual constraint on LogOdds(AandB).

Parameter	num_tokens	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0B	0.223	0.159	-0.087	0.539
Culture	0B	0.149	0.244	-0.327	0.623
Power	0B	0.286	0.249	-0.207	0.777
Freq	0B	-0.070	0.082	-0.232	0.091
Len	0B	0.030	0.127	-0.220	0.278
Intercept	2B	-0.027	0.256	-0.529	0.478
Culture	2B	-0.399	0.390	-1.161	0.361
Power	2B	0.531	0.404	-0.250	1.334
Freq	2B	0.258	0.135	-0.008	0.524
Len	2B	0.492	0.212	0.076	0.909
Intercept	41B	0.179	0.229	-0.268	0.628
Culture	41B	-0.377	0.347	-1.065	0.305
Power	41B	1.290	0.373	0.568	2.037
Freq	41B	-0.035	0.121	-0.274	0.202
Len	41B	0.807	0.188	0.438	1.179

Table B.0.1.: Model results examining the effect of each individual constraint on LogOdds(AandB).

Parameter	num_tokens	Estimate	Est.Error	Q2.5	Q97.5
Intercept	209B	0.176	0.182	-0.186	0.537
Culture	209B	0.290	0.283	-0.268	0.847
Power	209B	0.760	0.289	0.194	1.327
Freq	209B	-0.056	0.096	-0.244	0.132
Len	209B	-0.063	0.150	-0.358	0.234
Intercept	419B	-0.458	0.183	-0.816	-0.099
Culture	419B	-0.125	0.282	-0.679	0.437
Power	419B	0.508	0.289	-0.056	1.073
Freq	419B	0.240	0.096	0.053	0.431
Len	419B	-0.298	0.150	-0.591	-0.005
Intercept	838B	-0.022	0.184	-0.381	0.335
Culture	838B	-0.111	0.284	-0.661	0.446
Power	838B	0.865	0.297	0.283	1.456
Freq	838B	0.127	0.099	-0.068	0.319
Len	838B	0.247	0.154	-0.055	0.552
Intercept	1677B	-0.181	0.176	-0.527	0.159
Culture	1677B	0.861	0.273	0.326	1.394
Power	1677B	0.562	0.275	0.031	1.108
Freq	1677B	0.052	0.091	-0.125	0.230
Len	1677B	-0.431	0.142	-0.708	-0.156

Appendix C.

Full List of Stimuli

Table C.0.1.: Full list of binomials as well as their constraints.

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Final Stress	AbsPref
kiwis	wolverines	0	0	0	-1	-1	0	-0.29	1	-1	-1	0.43
kiwis	narwhals	0	1	0	-1	-1	0	2.71	0	-1	-1	0.51
kiwis	ocelots	0	0	0	-1	0	0	2.74	0	-1	-1	0.46
ibex	kiwis	0	0	0	1	0	0	-1.17	0	1	0	0.50
harpies	kiwis	0	-1	0	1	1	0	-1.95	0	0	0	0.47
axolotls	wolverines	0	-1	0	-1	0	0	-2.69	-1	-1	-1	0.26
axolotls	ibex	0	-1	0	-1	0	0	-1.23	-2	-1	0	0.33
axolotls	harpies	0	1	0	-1	-1	0	-0.45	-2	0	0	0.41
axolotls	keas	0	0	1	-1	0	0	1.05	-3	-1	1	0.59
axolotls	bonobos	0	-1	0	-1	0	0	-0.89	-1	0	0	0.33
axolotls	wombats	0	-1	0	-1	0	0	-0.65	-2	-1	-1	0.27
axolotls	lions	0	-1	-1	-1	-1	0	-5.11	-2	0	0	0.16
ocelots	platypuses	0	1	-1	1	0	0	0.67	1	3	1	0.53
ibex	platypuses	0	1	0	1	0	0	2.23	2	3	1	0.69
harpies	platypuses	0	0	0	1	1	0	1.45	2	2	0	0.58
keas	platypuses	0	0	-1	0	0	0	-0.05	3	3	1	0.46
bonobos	platypuses	0	1	0	1	0	0	1.90	1	2	0	0.61
harpies	wolverines	0	-1	0	0	0	0	-2.24	1	-1	1	0.57
keas	wolverines	0	-1	-1	-1	0	0	-3.74	2	0	0	0.28
bonobos	wolverines	0	0	0	0	0	0	-1.79	0	-1	-1	0.43
capybaras	ibex	0	0	0	0	0	0	-1.66	-2	-1	-1	0.36
capybaras	harpies	0	1	0	-1	-1	0	-0.88	-2	0	0	0.40
capybaras	keas	0	1	1	1	0	0	0.62	-3	-1	-1	0.59
ibex	narwhals	0	1	0	-1	0	0	1.54	0	0	0	0.54
harpies	narwhals	0	-1	0	0	1	0	0.77	0	-1	-1	0.42
keas	narwhals	0	0	-1	-1	0	0	-0.74	1	0	0	0.36
ibex	ocelots	0	0	0	0	0	0	1.57	1	0	0	0.58
keas	ocelots	0	-1	-1	-1	0	0	-0.71	2	0	0	0.34
bonobos	ocelots	0	0	1	0	0	0	1.23	0	-1	-1	0.59
harpies	ibex	0	-1	0	0	1	0	-0.78	0	-1	-1	0.39
ibex	keas	0	1	1	1	0	0	2.28	-1	0	0	0.73
bonobos	ibex	0	0	0	0	0	0	-0.33	-1	-1	-1	0.42
ibex	koalas	0	0	-1	0	0	0	-0.43	1	1	1	0.47
ibex	sloths	0	0	-1	1	0	0	-0.04	-1	0	0	0.43

Table C.0.1.: Full list of binomials as well as their constraints.

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Final Stress	AbsPref
aardvarks	ibex	0	0	1	0	0	0	-1.94	0	0	0	0.57
harpies	keas	0	-1	1	1	1	0	1.50	-1	-1	-1	0.57
bonobos	harpies	0	1	0	-1	-1	0	0.44	-1	0	0	0.47
harpies	koalas	0	-1	-1	0	1	0	-1.21	1	0	0	0.36
harpies	wombats	0	-1	0	1	1	0	-0.20	0	-1	-1	0.47
aardvarks	harpies	0	1	0	0	-1	0	-1.17	0	1	1	0.58
bonobos	keas	0	1	1	1	0	0	1.95	-2	-1	-1	0.66
keas	koalas	0	-1	-1	-1	0	0	-2.71	2	1	1	0.34
keas	sloths	0	-1	-1	0	0	0	-2.31	0	0	0	0.30
keas	wombats	0	-1	-1	-1	0	0	-1.71	1	0	0	0.29
keas	lions	0	-1	-1	-1	-1	0	-6.17	0	1	1	0.22
aardvarks	keas	0	1	1	1	0	0	0.34	-1	0	0	0.70
bonobos	wombats	0	0	0	1	0	0	0.24	-1	-1	-1	0.50
aardvarks	bonobos	0	0	0	0	0	0	-1.61	1	1	1	0.55
ocarinas	vibraphones	0	0	1	0	0	0	0.46	-1	-1	-1	0.54
cymbals	ocarinas	0	0	1	1	0	0	3.48	2	0	0	0.79
clarinets	ocarinas	0	0	1	1	0	0	2.24	0	1	1	0.74
cellos	ocarinas	0	0	1	0	0	0	2.34	2	0	0	0.72
didgeridoos	vibraphones	0	0	0	0	0	0	0.53	-1	0	0	0.48
lutes	marimbas	0	0	0	0	0	0	1.22	2	1	1	0.65
kalimbas	lutes	0	0	0	0	0	0	-2.41	-2	-1	-1	0.34
clarinets	kalimbas	0	0	1	1	0	0	2.83	0	1	1	0.75
kalimbas	trumpets	0	0	-1	-1	0	0	-4.44	-1	0	0	0.23
cellos	kalimbas	0	0	1	1	0	0	2.93	1	0	0	0.75
kalimbas	saxophones	0	0	-1	-1	0	0	-3.12	0	-1	-1	0.25
lutes	saxophones	0	0	-1	-1	0	0	-0.72	2	0	0	0.40
casserole	eagle	0	-1	0	0	-1	0	-1.53	-1	1	1	0.41
kite	linguist	0	-1	1	0	0	0	1.60	1	1	1	0.66
algorithm	perfume	0	-1	-1	0	0	0	1.99	-2	-1	-1	0.28
forest	screwdriver	0	0	0	0	0	0	3.29	1	0	0	0.61
slipper	volcano	0	0	1	-1	-1	0	-1.81	1	0	0	0.54
harmonica	microscope	0	0	0	0	0	0	-1.75	-1	-2	-1	0.44
cookbook	zenith	0	1	1	0	0	0	1.14	0	1	1	0.72
hammock	hydrogen	0	1	-1	0	0	0	-2.06	1	1	0	0.41
neuron	toaster	0	-1	-1	0	0	0	0.38	0	1	0	0.31
marshmallow	telescope	0	0	0	0	0	0	-1.17	0	-1	-1	0.44
casserole	optics	0	1	0	0	0	0	-0.66	-1	1	1	0.56
encyclopedia	comet	0	1	0	-1	0	0	0.95	-4	-1	0	0.42
nimbus	waffle	0	-1	-1	0	0	0	-1.68	0	0	0	0.31
photon	pumpkin	0	-1	-1	0	0	0	-0.79	0	1	1	0.37
lantern	syntax	0	1	1	0	0	0	-0.94	0	-1	-1	0.61
echo	vineyard	0	-1	0	0	0	0	1.88	0	0	0	0.48
nebula	snowman	0	-1	-1	0	0	0	0.18	-1	-2	-1	0.32
botany	teapot	0	-1	-1	0	0	0	0.44	-1	-2	-1	0.32
chisel	kaleidoscope	0	0	0	1	0	0	0.14	2	-1	-1	0.61
lava	teacup	0	-1	-1	1	0	0	1.97	0	-1	-1	0.41
entropy	orchard	0	-1	-1	0	0	0	0.37	-1	-1	0	0.36
axolotl	vineyard	0	1	0	0	0	0	-3.51	-2	0	0	0.42
clockwork	meadow	0	-1	0	0	0	0	-0.99	0	1	1	0.46
algebra	telescope	0	-1	0	0	0	0	8.75	0	-2	-1	0.63
arcade	topaz	0	0	0	0	0	0	2.28	0	0	0	0.55
asteroid	compass	0	-1	0	1	0	0	-0.86	-1	0	0	0.45

Table C.0.1.: Full list of binomials as well as their constraints.

Word1	Word2	Form	Percept	Culture	Power	Intense	Icon	Freq	Len	Lapse	Final Stress	AbsPref
bicycle	nebula	0	1	1	0	0	0	1.99	0	0	0	0.70
bungalow	entropy	0	1	0	0	0	0	-1.20	0	2	1	0.54
carnation	gnome	0	0	0	0	0	0	-1.77	-2	-1	-1	0.35
cinnamon	harmonica	0	0	1	0	0	0	2.30	1	0	0	0.69
coral	syntax	0	1	1	0	0	0	-0.02	0	-1	-1	0.63
dandelion	pendulum	0	0	0	0	0	0	-0.53	-1	1	0	0.41
delirium	telescope	0	-1	-1	0	1	0	-1.44	-1	-2	-1	0.29
anchors	sandstorms	0	1	0	0	-1	0	3.42	0	-1	-1	0.59
scissors	volcanoes	0	0	1	-1	0	0	0.58	1	0	0	0.59
equations	lanterns	0	-1	0	0	0	0	2.30	-1	0	0	0.45
satellites	tulips	0	-1	0	0	0	0	1.49	-1	1	1	0.48
compasses	hedgehogs	0	-1	0	0	0	0	-0.61	-1	-2	-1	0.40
comets	neckties	0	-1	0	1	0	0	2.29	0	-1	-1	0.52
castles	headphones	0	0	-1	1	0	0	-1.10	0	-1	-1	0.40
paperclips	pyramids	0	0	1	-1	0	0	-2.88	0	0	0	0.48
constellations	kettles	0	-1	0	0	0	0	0.94	-2	0	0	0.39
kaleidoscopes	whales	0	-1	-1	-1	0	0	-4.65	-3	0	0	0.15
meadows	pianos	0	-1	-1	0	0	0	1.15	1	0	0	0.40
magnets	zebras	0	-1	1	0	0	0	1.82	0	0	0	0.59
parrots	submarines	0	1	0	-1	0	0	-0.35	1	-1	-1	0.49
crayons	jungles	0	1	1	0	0	0	0.29	0	0	0	0.67
harbor	teapot	0	-1	0	0	0	0	2.56	0	-1	-1	0.46
notebook	quicksand	0	0	1	-1	0	0	3.31	0	0	0	0.61
glacier	lantern	0	-1	0	0	0	0	0.29	0	0	0	0.45
microscope	puddle	0	0	0	0	0	0	1.36	-1	1	1	0.54
compass	swan	0	-1	0	0	0	0	0.20	-1	-1	-1	0.37
bonsai	cathedral	0	1	0	-1	0	0	-2.00	1	1	1	0.54
honeycomb	violin	0	0	-1	0	0	0	-1.40	0	0	0	0.37
sailboat	stadium	0	0	0	0	0	0	-3.21	1	2	1	0.47
acorns	skyscrapers	0	1	0	0	0	0	-0.40	1	1	1	0.64
bell	trellis	0	0	1	0	0	0	3.33	1	1	1	0.74
inkwell	kite	0	0	-1	0	0	0	-3.22	-1	0	0	0.30
foxglove	trombone	0	0	-1	0	0	0	-1.89	0	1	1	0.40
carousel	quill	0	0	1	0	0	0	0.94	-2	0	0	0.55
lighthouse	onion	0	-1	-1	0	0	0	-1.15	0	1	1	0.36
cactus	chessboard	0	0	0	0	1	0	2.08	0	-1	-1	0.51
gallery	raindrop	0	0	0	0	0	0	5.06	-1	-2	-1	0.58
cricket	plow	0	1	0	-1	0	0	2.37	-1	-1	-1	0.47
gingerbread	fresco	0	0	1	0	0	0	0.35	-1	1	1	0.62
cello	sunflower	0	0	0	0	0	0	-0.42	1	0	0	0.53
archway	quilt	0	0	0	0	0	0	-2.78	-1	0	0	0.41
compass	haystack	0	0	0	0	0	0	2.34	0	-1	-1	0.51
beacon	millipede	0	-1	0	0	1	0	4.03	1	-1	-1	0.53
parchment	windmill	0	0	0	0	0	0	0.99	0	-1	-1	0.49
candlestick	meadow	0	1	0	0	0	0	-1.46	-1	1	1	0.54