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

Zachary Houghton[1], Jungah Lee[2], Casey Felton[1], Georgia Zellou[1], & Emily Morgan[1]

[1] University of California, Davis

[2] Chosun University

Author Note

Correspondence concerning this article should be addressed to Zachary Houghton.
E-mail: znhoughton@ucdavis.edu

Abstract

The question of what items are stored in the lexicon is one that has drawn a lot of attention in the last few decades, and while the general consensus is that a lot more is stored than we previously realized, it is still largely unclear what factors drive storage. For example, some have argued that frequency drives storage, while others have posited that predictability drives storage. Further, it is unclear what the relationship between stored multi-word items and the representation of each individual word is. For example, it is possible that stored items fuse together, losing some amount of their internal structure. The present paper examines both of these questions by looking at the recognizability of the segment *up* in English V+*up* phrases. We find that the time it takes to recognize *up* decreases as frequency or predictability increases, but increases once again for the highest frequency or highest predictability items. Our results suggest that frequency and predictability both drive storage, and that stored items may lose some amount of their internal representation.

Keywords: Psycholinguistics, holistic storage, language processing, lexical processing, phonological processing

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

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 (e.g., Bybee, 2002, 2003; Goldberg, 2003; Nooteboom, Nooteboom, Weerman, & Wijnen, 2002; Stemberger & MacWhinney, 1986, 2004).

Throughout the years different theories have argued for different degrees of holistic storage. Two theories in particular have dominated the field. On one hand, Chomskyan theories (e.g., Chomsky, 1965) have proposed that only necessary items (e.g., items that can't be formed compositionally) are stored. On the other hand, usage-based theories (e.g., Bybee, 2003) have proposed that multi-word items can be stored under certain usage-based conditions, such as frequency of use.

Traditional Chomskyan theories (e.g., Chomsky, 1965) 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 expressions 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, Liu, & Wang, 2003). This is magnitudes larger than any current estimate of how much storage language requires[1]. Considering this, it might not come as a surprise that there has been a rise in support for usage-based theories over the past few decades (Ambridge, 2020; Baayen, Schreuder, De Jong, & Krott, 2002; Bybee, 2002, 2003; Bybee & Hopper, 2001; Bybee & Scheibman, 1999; Kapatsinski, 2018; Kapatsinski & Radicke, 2009; Morgan & Levy, 2016; Stemberger & MacWhinney, 1986, 2004; Zang, Wang, Bai, Yan, & Liversedge, 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-compositional items, multi-word phrases such as *I don't know* may also be stored holistically if they are used frequently enough (Ambridge, 2020; Arnon & Snider, 2010; Hay, 2001; e.g., Kapatsinski, 2018; Kapatsinski & Radicke, 2009; Lee & Kapatsinski, 2015; Morgan & Levy, 2016; Stemberger & MacWhinney, 1986, 2004; Tomasello, 2005).

While it has become a dominant view in the field that at least some multi-word items

[1] Indeed, Mollica and 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.

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?

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; Hay, 2001; Stemberger & MacWhinney, 1986, 2004; Zwitserlood, 2018), especially in the phonology literature. For example, Bybee and Scheibman (1999) demonstrated that the word *don't* is reduced to a larger extent in the phrase *I don't know* than in other words 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. Additionally, 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 phonological effects, the Psycholinguistics literature has also provided an abundance of evidence for multi-word storage. For example, Siyanova-Chanturia, Conklin, and Heuven (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 and Levy (2016) 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), providing additional evidence that frequent binomials are stored holistically.

Further, there is also evidence of multi-word storage from the learning literature. For example, Siegelman and 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 much 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. Following their results it seems reasonable that storing items holistically may facilitate learning.

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 (Bybee & Scheibman, 1999; Kapatsinski & Radicke, 2009; Lee & Kapatsinski, 2015; e.g., Maye & Gerken, 2000) and predictability (e.g, Olejarczuk, Kapatsinski, & Baayen, 2018; Ramscar, Dye, & Klein, 2013).

Traditionally, frequency has been assumed to be the driving factor behind multi-word storage, and for good reason. 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, the fact that sound changes occur differently depending on the frequency of the word/phrase suggests that they have separate representations (i.e., holistic storage).

On the other hand, predictability has also been shown to play a crucial role in learning (Olejarczuk et al., 2018; Ramscar et al., 2013; Saffran, Aslin, & Newport, 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 events, suggesting that the learning of phonetic categories is also driven by prediction. 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 connection may be clear by now: 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 given the abundance of evidence for both frequency and predictability effects in language, it remains unclear to what extent each of these factors drives storage.

1.3 Processing Consequences of Storage

Given the evidence that a lot more may be stored than previously thought, another important question to consider is what exactly the processing consequences of storage are. 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.* On the other hand, as the phrase is used more often, it is possible that it loses some amount of its internal representation with respect to its component parts.

Indeed, there seems to be some evidence that multi-word phrases may lose some amount of their internal structure. For example, Kapatsinski and 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 higher frequency phrasal verbs they were quicker to recognize *up*. However, upon reaching the highest frequency words participants grew slower to recognize *up* Kapatsinski & Radicke (2009). 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 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.

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 absent in the holistic representation.

Additionally, there is also evidence from the word-recognition literature that some stored words may also lose some of their internal structure as well. For example, Healy (1976) examined participants' ability to recognize letters in various words. He found that people were worse at recognizing the letter *t* in *the* than in other lower frequency words,

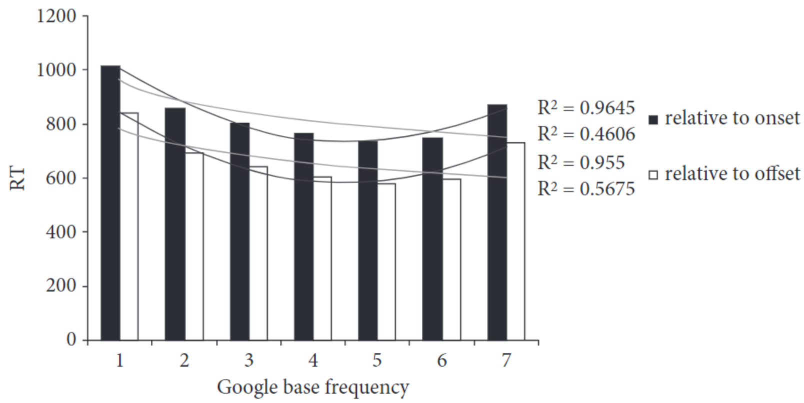

Figure 1. The U-shaped effect of the frequency of verb+*up* constructions on the speed with which up is detected, reproduced from Kapatsinski and Radicke (2009).

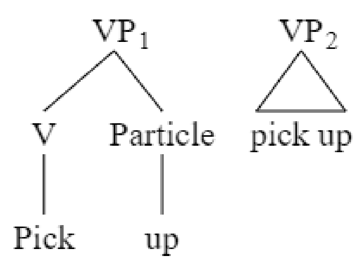

Figure 2. A diagram of two ways the word *pick up* could be stored. The left tree demonstrates a stored representation of *pick up*, where the internal structure is still intact. The right tree demonstrates a holistically stored unit, where there is a loss of internal structure. Note that these are stored structures, as opposed to a compositional representation of *pick up* which would be comprised of the individual representations *pick* and *up*.

which suggests that even words can develop a representation separate from its component pieces (in this case, the component parts being letters instead of words). If it is the case that *the* is recognized as a composition of its parts, then it's unclear how to account for these results (c.f., Kapatsinski & Radicke, 2009, who suggested that one explanation is that people don't fixate as long on high-frequency and function words, of which *the* is both).

On the other hand, there is a necessary temporal linearity to speech, so listeners recieve the information for some of the component parts before the entire phrase. For example, the listener hears *pick* before *pick up*. It seems a bit unlikely that a listener would process *pick up* without having processed *pick* at all. Thus if holistically stored phrases do lose the representation of their component parts, it's unclear what exactly the relationship with processing is. One such possibility that was put forth by both Kapatsinski and Radicke (2009) and Healy (1976) is that during processing, the holistic representations compete with the representations of the individual parts for recognition. In other words: once listeners finish processing the phrase, they move on, even if they haven't processed the individual parts.[2] This is necessary to account for the results in Kapatsinski and Radicke (2009) because high-frequency phrases are still processed more quickly than lower frequency phrases. If accessing the holistic representation facilitates the accessing of the individual parts (as James L. McClelland & Rumelhart, 1981's model suggests), then we would see a decrease in recognition times for the parts. However, an increase in recognition times suggests there is competition for recognition between the holistic representation and the representations of the individual parts.

1.4 Present Study

The present study examines the factors that drive storage and the processing consequences of storage by extending Kapatsinski and Radicke (2009) to look at the effects

[2] Note that competition can be implemented in other ways though, e.g., using top-down inhibition (Libben, 2005).

of both frequency, predictability, and their interaction on the processing of V+*up* phrases. Similar to Kapatsinski and 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 both frequency and predictability. Since frequency effects are rather robust in the literature, we should at the very least see a negative correlation between frequency and reaction time (up to perhaps a certain point, where recognition may get harder). The effects of predictability on recognition times, however, are still relatively untested in the literature. If predictability is not a driving factor of storage, we should see only frequency effects on the recognizability of *up*. Further, if storage does result in a loss of internal structure, we should see similar effects to those found in Kapatsinski and Radicke (2009). Specifically, we should see some sort of U-shaped effect, where recognition gets easier until we get to the highest frequency/predictability items, where recognizability should then become harder.

2 Methods

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 SONA 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.

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 (1)):

$$\frac{\text{count}(\mathit{Verb}+\mathit{up})}{\text{count}(\mathit{Verb}) - \text{count}(\mathit{Verb}+\mathit{up})} \tag{1}$$

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 aimed to gather 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 below.

Some verb phrases containing *up* display unique syntactic patterns. For example, see the below phrasal verbs:

(1) a. The controversy stirred up a heated debate.

b. ??The controversy stirred a heated debate up.

These verbs show a syntactic alternation that is not present in all verb+*up* collocations (e.g., *stirred up a heated debate* is fine, but *stirred a heated debate up* is weird

[3] Levy, Fedorenko, Breen, and Gibson (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.

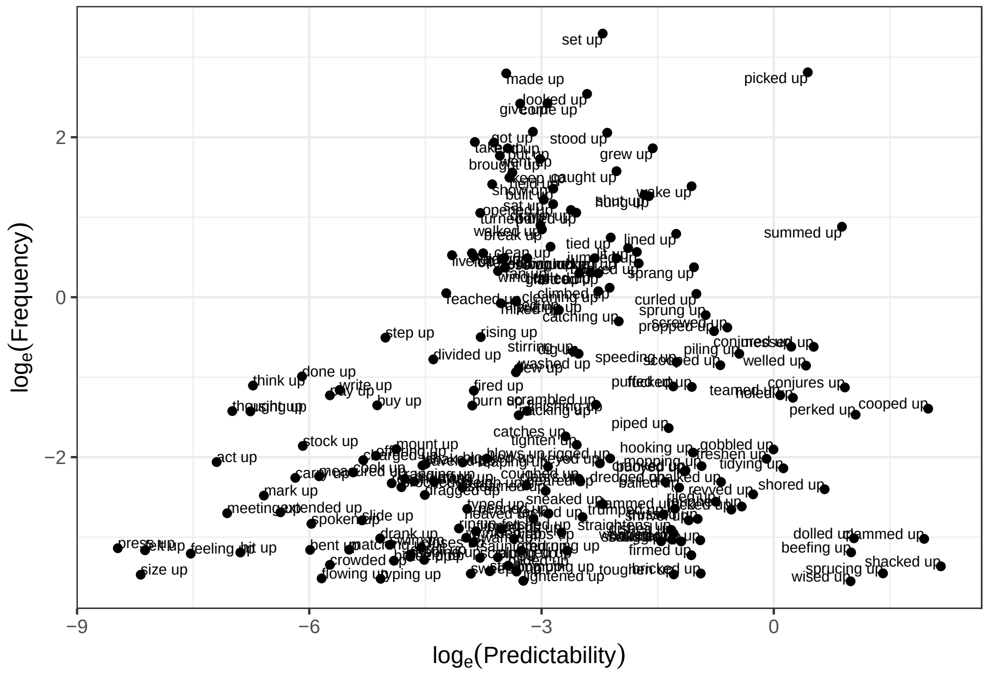

Figure 3. log-predictability by log-frequency (per million) plot of our items.

at best). It is possible that due to this syntactic alternation, phrasal verbs may be stored regardless of frequency/predictability. 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.

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 match the frequency range of our experimental stimuli, we extracted the 50 most frequent words, then sampled from the rest of the dataset to gather an additional 100 words. These 350 items together comprise our stimuli.

For each item, we constructed two sentences: one sentence which contained *up*, and one sentence that was identical except that it didn't include the segment *up.* For words, the entire word was replaced. For phrases, *up* was simply deleted if possible (e.g., *clean up* replaced with *clean*). If resulting 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 with a different morpheme or word.

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.

2.3 Procedure

Participants were presented with audio sentences via Pavlovia (https://pavlovia.org/), a website for presenting PsychoPy experiments. 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.

2.4 Analysis

The data[4] was analyzed using General Additive Mixed models, as implemented in the *mgcv* package (Wood, 2011) within the R programming environment (R Core Team, 2023). 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).

For all of our models, the dependent variable was the time it took for participants to react to the onset of the target segment (i.e., the time it took participants to press the button after hearing *up*). For the first model, the predictors were 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, and trial. Our model formula is included below in Equation (2):

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

$$
\begin{aligned}
log(RT) \sim\ & ti(Predictability, Frequency) + Duration + s(participant, bs =\text{`} re\text{'}) + s(Item, bs =\text{`} re\text{'}) \\
& + s(trial, bs =\text{`} re\text{'}) + s(Predictability, Frequency, participant, bs =\text{`} re\text{'})
\end{aligned} \tag{2}
$$

We also ran an additional analysis similar to the first model, but allowing the interaction to vary for phrasal vs non-phrasal verbs. Specifically, the model is identical to the first model with the exception that the effect of the interaction term was allowed to be different for phrasal verbs and non-phrasal verbs. This was done in order to examine whether the effect of frequency and predictability was different for phrasal verbs versus non-phrasal verbs. See Equation (3):

$$
\begin{aligned}
log(RT) \sim\ & ti(Predictability, Frequency, by = PhrasalVerb) + Duration + s(participant, bs =\text{`} re\text{'}) \\
& + s(Item, bs =\text{`} re\text{'}) + s(trial, bs =\text{`} re\text{'}) + s(Predictability, Frequency, participant, bs =\text{`} re\text{'})
\end{aligned} \tag{3}
$$

Additionally, we ran a Generalized Additive Model with frequency, predictability, and the interaction between frequency and predictability as fixed-effects that could vary non-linearly, and duration of the segment 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 (4):

$$
\begin{aligned}
log(RT) \sim\ & s(Predictability) + s(Frequency) + ti(Predictability, Frequency) + Duration \\
& + s(participant, bs =\text{`} re\text{'}) + s(Item, bs =\text{`} re\text{'}) + s(trial, bs =\text{`} re\text{'}) \\
& + s(Predictability, Frequency, Trial, Participant, bs =\text{`} re\text{'})
\end{aligned} \tag{4}
$$

Finally, we replicated the analyses from Kapatsinski and 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 model syntax for both models is included below in Equations (5) and (6):

$$
\begin{aligned}
log(RT) \sim\ & log(Frequency) + Duration + log(Frequency^2) \\
& + (1 + log(Frequency) + log(Frequency^2) + Duration||Participant) + (1||Item)
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
log(RT) \sim\ & log(Predictability) + Duration + log(Predictability^2) \\
& + (1 + log(Predictability) + log(Predictability^2) + Duration||Participant) + (1||Item)
\end{aligned} \tag{6}
$$

3 Results

The effect of the interaction between frequency and predictability was not significant in any of our models (see Tables 1 through 3 for the output of each model). Further, there was no significant effect of whether the verb phrase was a phrasal verb (e.g., *pick up*) or not (e.g., *walk up*)[5]. In other words, the effects are the same regardless of whether the item was

[5] A BIC analysis confirmed that the model that included whether the verb phrase was a phrasal verb or not (analysis in Table 2) was not a better fit than the identical model without it (the analysis in Table 1).

a phrasal verb or not. Additionally, our third Generalized Additive Model suggested that there was a significant main-effect of predictability.

Table 1

Model results for the Generalized Additive Mixed Model containing 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 2

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

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

Table 3

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

	edf	Ref.df	F	p-value
s(log-frequency)	2.43	2.48	1.68	0.320
s(log-predictability)	1.88	1.92	3.30	0.030
s(log-frequency*log-predictability)	0.00	0.00	0.05	0.990
s(participant)	296.33	305.00	37.57	<0.001
s(item)	176.42	196.00	10.64	<0.001
s(log-pred., log-freq., log-freq.:log-pred., participant)	0.02	306.00	0.00	0.750

Given these results, we ran a follow-up Bayesian quadratic regression model to further examine the effects. Since the Generalized Additive Model suggested that there was no significant interaction between frequency and predictability, we left out the interaction term from the regression model. We also modeled the random-effects without correlations between them (this was done to allow the model to run faster, since we collected a large amount of data). Equation (7) below presents the full model syntax:

$$
\begin{aligned}
log(RT) \sim\ & log(Frequency) + log(Predictability) + Duration + log(Frequency^2) + log(Predictability^2) \\
& + (1 + log(Frequency) + log(Predictability) + log(Frequency^2) + log(Predictability^2) \\
& + Duration||Participant) + (1||Item)
\end{aligned} \tag{7}
$$

The results of this model are presented below in Table 4. Following Houghton, Kato, Baese-Berk, and Vaughn (2023), in cases where the confidence interval crosses zero, we also report the percentage of posterior samples greater than or less than zero. For the current model, although the confidence 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 4.

Table 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
Intercept	-0.10	0.03	-0.16	-0.05
log-frequency	0.02	0.01	0.00	0.04
log-predictability	0.01	0.01	-0.01	0.03
duration	-0.14	0.10	-0.33	0.06
log-predictability^2	0.00	0.00	0.00	0.01
log-frequency^2	0.00	0.00	0.00	0.01

Finally tables 5 and 6 present the results for the quadratic regression models including only frequency and frequency^2 as well as the quadratic regression model including only predictability and predictability^2 respectively:

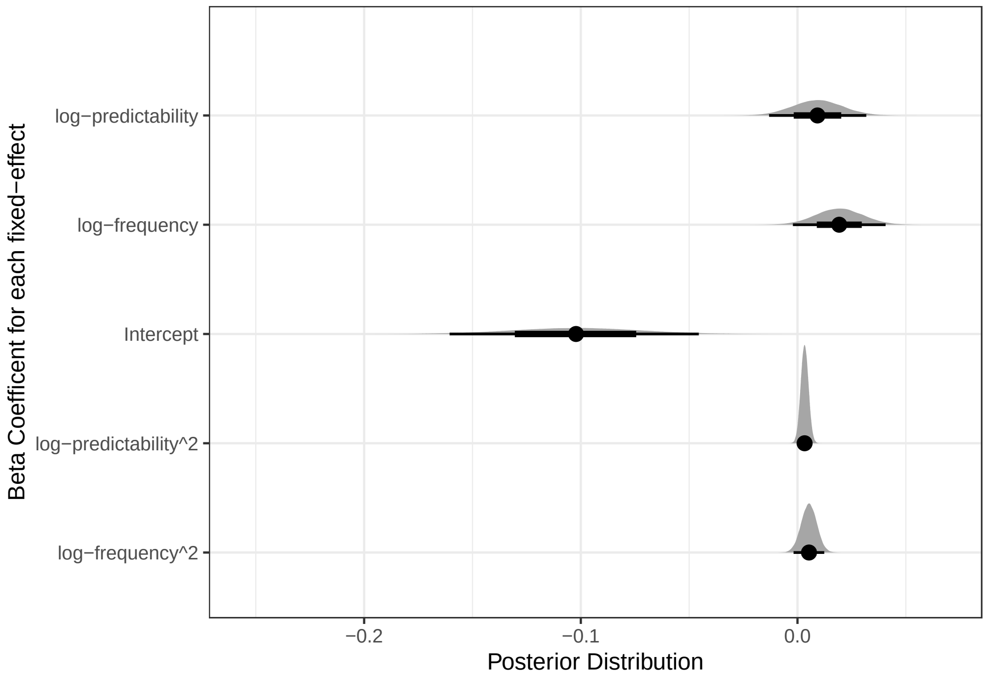

Figure 4. Plot of the posterior distribution for the beta value of each fixed-effect in our Bayesian quadratic regression model. The y axis contains the different fixed-effects and the x-axis contains the posterior distribution of beta values for the corresponding fixed-effect.

Table 5

Results for the Bayesian quadratic regression model containing only frequency and frequency^2.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	-0.10	0.02	-0.15	-0.05
log-frequency	0.02	0.01	0.00	0.04
Duration	-0.08	0.10	-0.27	0.11
log-frequency$\hat{}$2	0.01	0.00	0.00	0.01

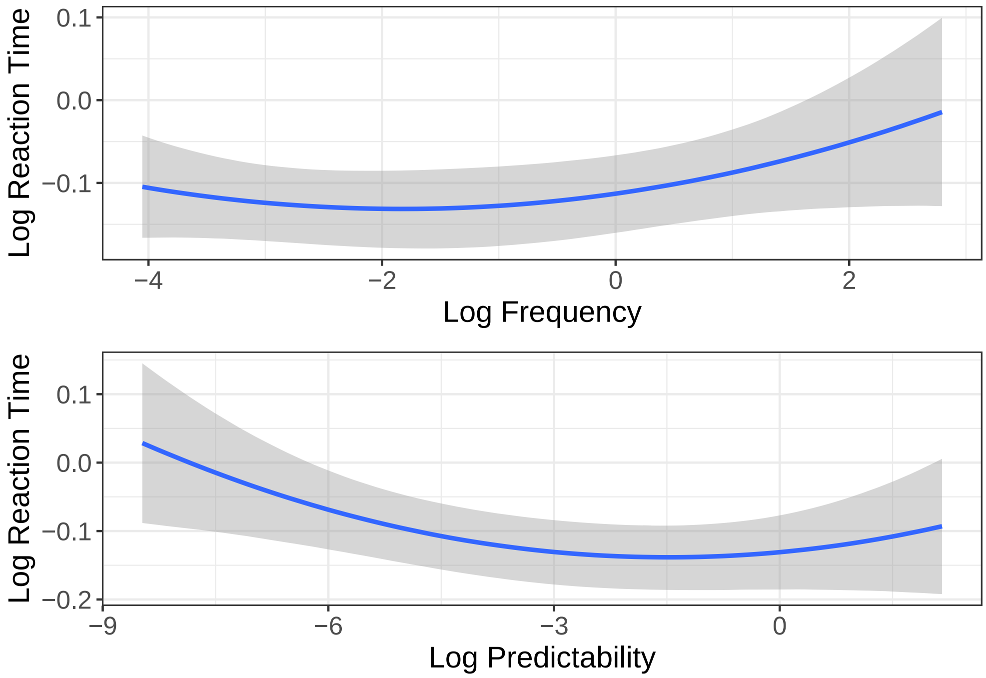

Figure 5. Visualization of the model results from Table 1 for frequency (top) and predictability (bottom). Frequencies are per million.

Table 6

Results for the Bayesian quadratic regression model containing only prediductability and predictability^2.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	-0.10	0.02	-0.15	-0.05
log-predictability	0.02	0.01	0.00	0.04
Duration	-0.08	0.10	-0.27	0.11
log-predictability^2	0.01	0.00	0.00	0.01

While the confidence interval for the quadratic term in both models crosses zero, over 95% of the posterior samples for log-frequency^2 were greater than zero and over 96 percent of the posterior samples for log-predictability^2 were greater than zero. A visualization of the posterior distributions for both models are presented in Figure 6 and Figure 7. Further, visualizations of the model predictions are also included below in Figures 8 and 9.

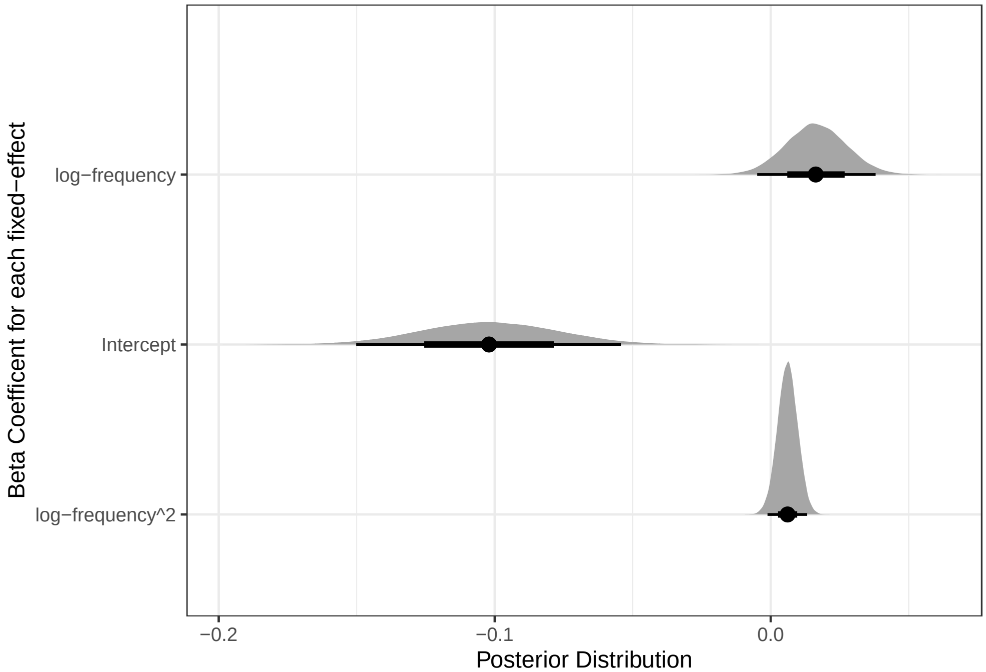

Figure 6. Plot of the posterior distribution for the beta value of each fixed-effect in our frequency-only quadratic regression model. The y axis contains the different fixed-effects and the x-axis contains the posterior distribution of beta values for the corresponding fixed-effect.

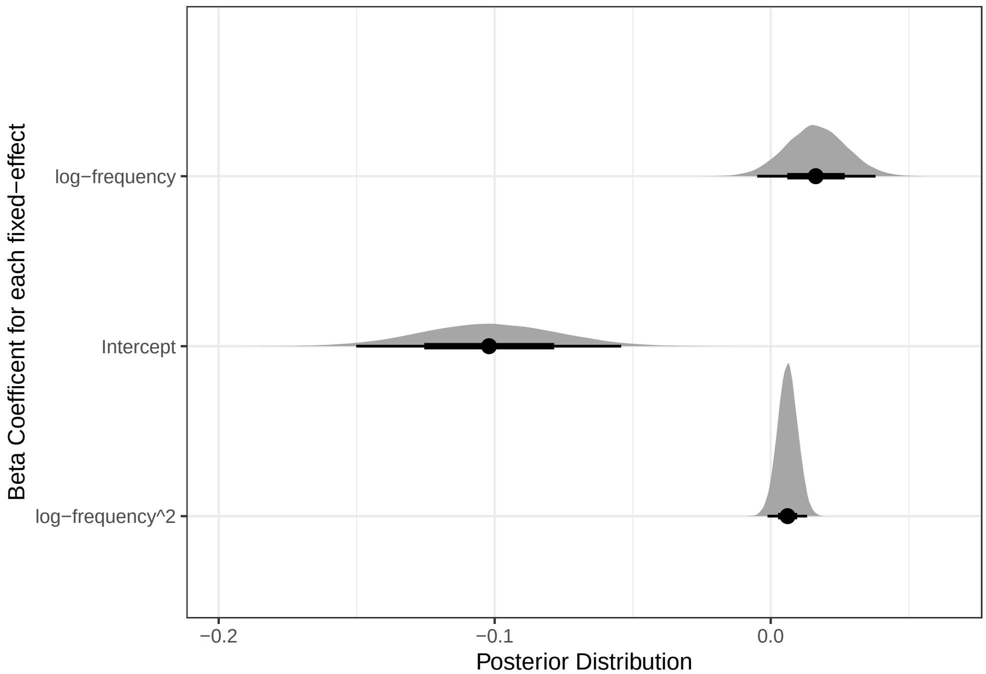

Figure 7. Plot of the posterior distribution for the beta value of each fixed-effect in our predictability-only quadratic regression model. The y axis contains the different fixed-effects and the x-axis contains the posterior distribution of beta values for the corresponding fixed-effect.

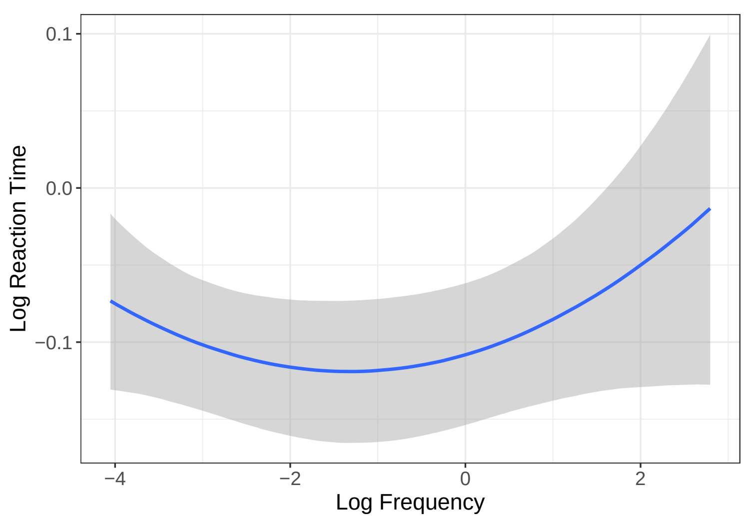

Figure 8. Model predictions for the effects of frequency on reaction times for the frequency-only Bayesian quadratic model.

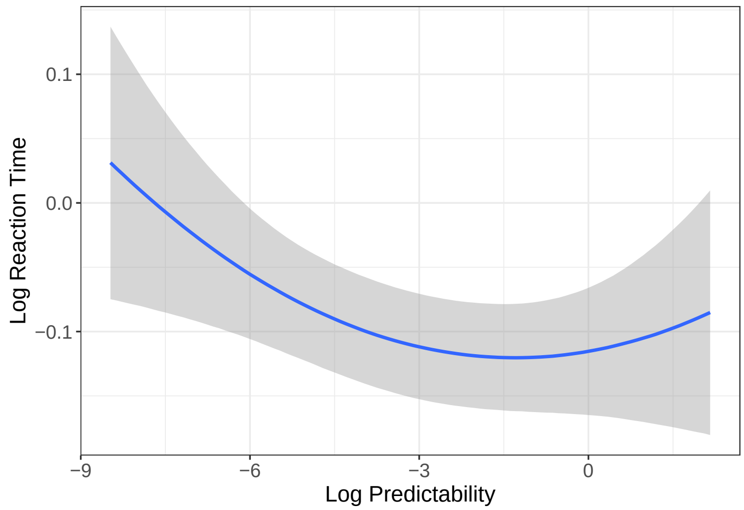

Figure 9. Model predictions for the effect of predictability on reaction times for the predictability-only models.

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/predictability increased, people were faster at recognizing *up*, until reaching the highest frequency/most predictable items, where people were slower. These results suggest that the most predictable and/or highest frequency items have a lack of internal structure. We also found no meaningful differences between phrasal verbs (e.g., *pick up*) and non-phrasal verbs (e.g., *walk up*), suggesting that this effect is driven primarily by the statistical distribution of the input as opposed to syntactic properties. Although future research would do well to examine whether semantic properties (such as the extent to which a phrase is semantically transparent) mediate this effect.

First, our results suggest that both frequency and predictability drive storage, as we see an increase in recognition times for the highest frequency and highest predictability items. It is unclear how this result can arise without storage, since in order for *up* to be harder to recognize in some contexts, it must have a separate representation.

Our results also show that as frequency or predictability increases, recognition time decreases until reaching the highest frequency/predictability items where there is an increase in recognition time. These findings demonstrate competition between different levels (e.g., competition between a holistically stored phrase and its individual parts) during processing. These results replicate previous findings (Healy, 1976, 1994; e.g., Kapatsinski & Radicke, 2009; Minkoff & Raney, 2000). For example, as stated earlier, Healy (1976) found that people make more letter-detection errors in high-frequency words (e.g., *the*) than in lower frequency words. Further, Minkoff and Raney (2000) found that letters are more difficult to detect in high-frequency nouns than in low-frequency nouns. These results suggest that recognizing words or holistically stored phrases does not necessarily require recognizing the individual letters or sounds that comprise them.

However, competition alone cannot explain the increase in recognition time. A high-frequency holistically stored representation with intact internal structure would show a similar decrease in recognition time for its component parts. This is because accessing the holistically stored representation, if its internal representation is intact, entails accessing the representations of the individual parts. However, one possible explanation is that the increase in recognition time reflects a loss of internal structure over time. That is, it is possible that over time, more experience with the phrases results in a loss of the internal structure, or a weakening of the associations between the individual words and the meaning of the phrase (as demonstrated in Figure 2).

On the other hand, it's possible for high-frequency and high-predictability items, when accessing the first word, e.g., *pick*, the listener accesses the phrase immediately, before hearing *up*, and then continues on 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 be recovering the phonological representation from the semantic representation of the phrase, causing a delay in recognition time.

Another possible account is that perhaps the internal structure for the high-frequency and high-predictability items was never learned to begin with. For example, 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 most predictable phrases may not be separated out of the speech stream initially.

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

If it is the case that the internal structure for the phrase was never learned, it would explain why we see an increase in recognition times for *up*: as one encounters the phrase more often, the association between the holistic representation and the words/sounds in the phrase increases. Even after one learns that *pick up*, for example, is composed of two words, the holistic representation will still be more strongly associated with the phrase and continue to be activated.

Further, if the lack of internal representation is a function of our learning mechanisms, it may not be surprising that both predictability and frequency drive this lack of representation, since our brain employs both Hebbian (frequency-driven learning) and error-driven learning mechanisms (i.e., predictability-driven learning, Ashby, Ennis, & Spiering, 2007; Kapatsinski, 2018).

Additionally, we see the same U-shaped effect in both phrasal (e.g., *pick up*) and non-phrasal verbs (e.g., *walk 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. Thus this result, at a glance, seems to be problematic for the interpretation of our results. Mainly, if our results indicate storage then if all phrasal verbs are stored it contradicts our interpretation. However, while loss of internal representation indicates storage, storage does not necessarily indicate loss of representation. It is the combination of storage and usage that leads to loss of internal representation. Thus, the interpretation of our results holds regardless of whether phrasal verbs as a whole are stored holistically.

Finally, our results provide some insight into the nature of competition during processing. For example, as mentioned earlier, competition can be implemented in different ways. For example, Healy (1976) has 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. While this is plausible in reading, it seems much less likely in speech processing. Since listeners receive auditory signal in a continuous stream, listeners don't have an option analogous to skipping to the next word in reading. Our results are instead more compatible with a race model with inhibition, such as the TRACE model James L. McClelland & Elman (1986) .

In summary, we demonstrate that both frequency and predictability drive the holistic storage of phrasal verbs, and these holistically stored items compete with their component parts during lexical access. Further, we demonstrate that the highest frequency and most predictable items do not have a fully intact internal representation. Future work would do well to examine if stored items are learned without internal structure or if the internal structure is lost over time as a function of experience.

5 References

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

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(3), 632. Retrieved from https://psycnet.apa.org/journals/rev/114/3/632/

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.). Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-94-010-0355-1_3

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

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.

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.* Retrieved from https://ntrs.nasa.gov/api/citations/19670002070/downloads/19670002070.pdf

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*(3), 542–551. https://doi.org/10.1111/tops.12274

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

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*(376), 1041–1070. https://doi.org/10.1515/ling.2001.041

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. Retrieved from https://psycnet.apa.org/journals/xhp/2/2/235/

Healy, A. F. (1994). Letter detection: A window to unitization and other cognitive processes in reading text. *Psychonomic Bulletin & Review*, *1*(3), 333–344. https://doi.org/10.3758/BF03213975

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. (2023). Task-dependent consequences of disfluency in perception of native and non-native speech. *Applied Psycholinguistics*, 1–17. https://doi.org/10.1017/S0142716423000486

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

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

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

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

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

Libben, G. (2005). Everything is psycholinguistics: Material and methodological considerations in the study of compound processing. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, *50*(1-4), 267283. Retrieved from https://www.cambridge.org/core/journals/canadian-journal-of-linguistics-revue-canadienne-de-linguistique/article/everything-is-psycholinguistics-material-and-methodological-considerations-in-the-study-of-compound-processing/BF01DE531A8F8305E2A664711A7C4DB3

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

Maye, J., & Gerken, L. (2000). *Learning phonemes without minimal pairs*. *2*, 522533.

McClelland, James L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 186. Retrieved from https://www.sciencedirect.com/science/article/pii/0010028586900150

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*. Retrieved from https://apps.dtic.mil/sti/citations/ADA157550

McClelland, James 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. Retrieved from https://psycnet.apa.org/record/1981-31825-001

Minkoff, S. R. B., & Raney, G. E. (2000). Letter-Detection Errors in the Word The: Word Frequency Versus Syntactic Structure. *Scientific Studies of Reading*, *4*(1), 55–76. https://doi.org/10.1207/S1532799XSSR0401_5

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. (2016). Frequency-dependent regularization in iterated learning. *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)*, (2015).

Nooteboom, S., Nooteboom, S., Weerman, F., & Wijnen, F. (2002). *Storage and computation in the language faculty*. Springer Science & Business Media.

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

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

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

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

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

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

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

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

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

Wood, S. N. (2011). *Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models*. *73*, 3–36.

Yi, B.-W. (2002). Ŭmun hyŏnsanggwa pindo hyogwa [the effects of frequency and phonology]. *Han'gugŏhak*, *15*, 161–183.

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

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