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

Word count: 6411

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: holistic storage; sentence processing; frequency effects; predictability effects

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, 2003; 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, Generativist theories (e.g., Pinker, 1991; 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.

Traditional Generativist theories (e.g., Pinker, 1991; 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, Generativist 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,

[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.

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).

Generativist 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; Baayen et al., 2002; Bybee, 2003; Bybee & Hopper, 2001; Bybee & Scheibman, 1999; Kapatsinski, 2018; Kapatsinski & Radicke, 2009; Morgan & Levy, 2016; 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-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, 2016; 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 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; 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 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

[2] 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.

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), 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 and 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 and 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.

Finally, there is also evidence of multi-word storage from the learning literature (Bannard & Matthews, 2008; Siegelman & Arnon, 2015). 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 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. Moreover, storing larger-than-word chunks may possibly be facilitating the learning of grammatical gender in their study.

1.2 What Drives Storage?

Despite the evidence for 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). One such study that did examine the role of predictability in holistic storage was Z. N. Houghton and Morgan (2023). Z. N. Houghton and Morgan (2023) 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) a. Jimmy spread out the peanut butter. *high-predictability, plausible*
b. Jimmy picked up the peanut butter. *high-predictability, implausible*

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. Our results suggested that either predictability doesn't drive the holistic storage of compound nouns or that it doesn't facilitate processing in this manner.

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.

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 (e.g., *pick*

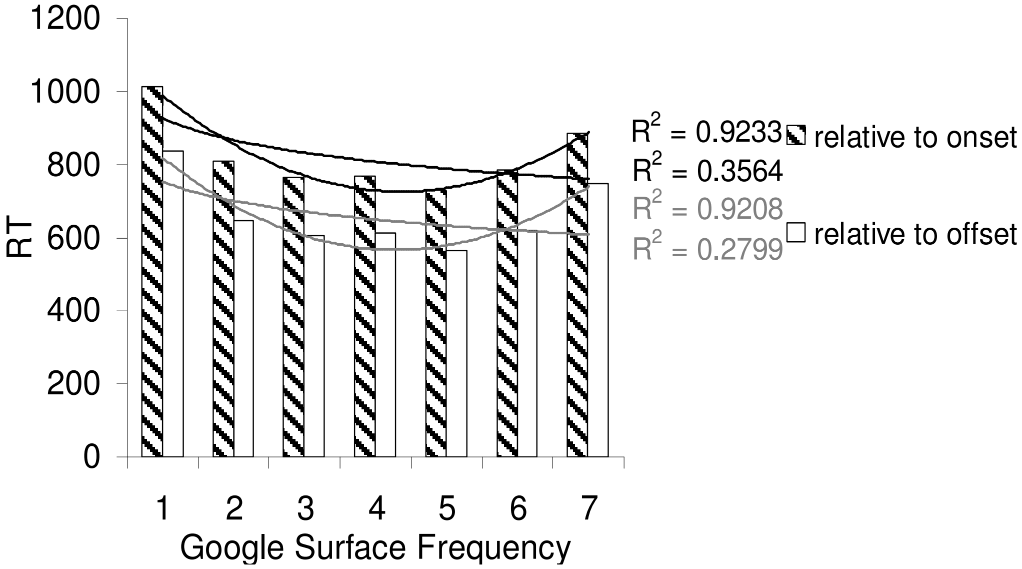

Figure 1: The U-shaped effect of the frequency of verb+*up* constructions on the speed with which *up* is detected, reproduced from Kapatsinski and Radicke (2009).

and *up*; see Figure 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 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 presented auditorily with 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 and for medium-high-frequency phrasal verbs they were quicker to recognize *up*. However, upon reaching the highest frequency words participants grew slower to recognize *up* (See Figure 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 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 and 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.

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

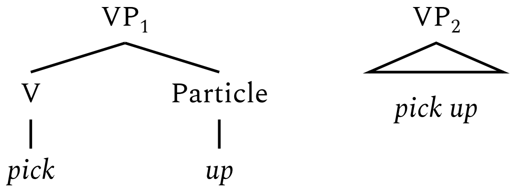

Figure 2: A diagram of two ways the word *pick up* could be stored. The left tree demonstrates a stored representation of *pick up*, where the internal structure is still intact. The right tree demonstrates a holistically stored unit, where there is a loss of internal structure. Note that both of these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

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.

1.4 Present Study

The present study examines the factors that drive storage and the representations of stored items by extending Kapatsinski and Radicke (2009) to look at the effects 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 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.

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 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.

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

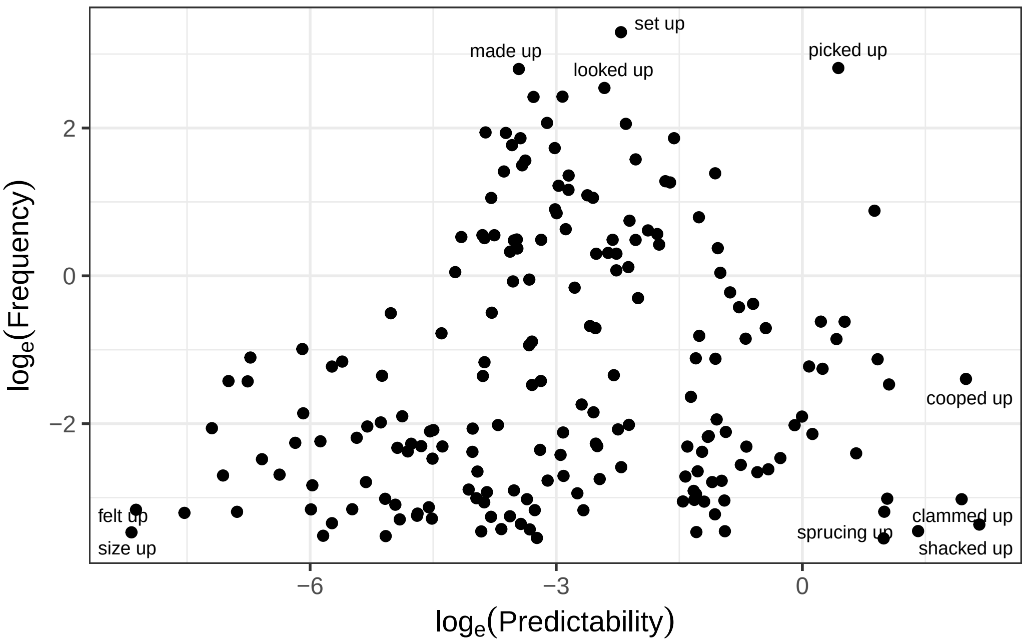

Figure 3: log-predictability by log-frequency (per million) plot of our items.

probability of *up* occurring immediately after the verb to the probability of any other word occurring (Equation 1).

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

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 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.

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 possible 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

[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.

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

alternate between the noun coming between the verb and the particle and the noun coming 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.

(2) 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 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.

(3) 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.

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. Results

The data 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, and trial. All our random-effects were allowed to be wiggly (non-linear). Our model formula is included below in Equation 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 the syntax `te()` indicates that the full tensor product smooth is used 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()`.

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

The results of this model are presented in Table 1 and visualized in Figure 4. We found no significant effect of the tensor product smooth.[5] 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:

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

Our results for this model are reported in Table 2 and visualized in Figure 5. Overall our results replicate the results from the the model that didn't include phrasal verb as a

[5] 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.

predictor (Equation 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 4:

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

Our results are presented in Table 3 and visualized in Figure 6. 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.[6]

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

In the Psycholinguistics literature, generalized additive mixed models are not yet well established. Thus, we ran a follow-up Bayesian quadratic regression model to further examine the effects of frequency and predictability on recognition times. Since the Generalized Additive Model suggested that there was no significant interaction between frequency and predictability, we left out the interaction term from the regression model. 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 5 below presents the full model syntax:

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

The results of this model are presented in Table 4 and visualized in Figure 7. Following Z. 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 8. 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

[6] 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.

the most frequent or most predictable items, where participants were slower to recognize *up*.

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 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 6 and Equation 7:

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

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

The results of our first model are presented in Table 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 5, $\beta = 0.005$).

The results of our second model are presented in Table 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 the effect from our full quadratic model model (Equation 5, $\beta = 0.003$). In other words, the results from both of our individual quadratic regression models (Equation 6 and Equation 7) replicate those found in Table 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 1: Model results for the generalized Additive Mixed Model cotanining only the interaction between frequency and predictability (Equation 2).

	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 (Equation 3).

	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 (Equation 4).

	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 4: Model results for the Bayesian quadratic regression model containing fixed-effects for frequency, predictability, and their quadratics (Equation 5).

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-0.10	0.03	-0.16	-0.05	0.03
log-frequency	0.02	0.01	0.00	0.04	96.16
log-predictability	0.01	0.01	-0.01	0.03	79.00
duration	-0.14	0.10	-0.33	0.06	8.27
log-predictability^2	0.00	0.00	0.00	0.01	96.88
log-frequency^2	0.00	0.00	0.00	0.01	92.94

Table 5: Results for the Bayesian quadratic regression model containing only frequency and frequency^2 (Equation 6).

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-0.10	0.03	-0.15	-0.05	0.00
log-frequency	0.02	0.01	0.00	0.04	93.31
Duration	-0.08	0.10	-0.27	0.11	19.36
log-frequency^2	0.01	0.00	0.00	0.01	95.23

Table 6: Results for the Bayesian quadratic regression model containing only predictability and predictability^2(Equation 7).

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-0.11	0.03	-0.16	-0.06	0.00
log-predictability	0.01	0.01	-0.01	0.03	75.74
Duration	-0.09	0.10	-0.28	0.10	18.42
log-predictability^2	0.00	0.00	0.00	0.01	96.10

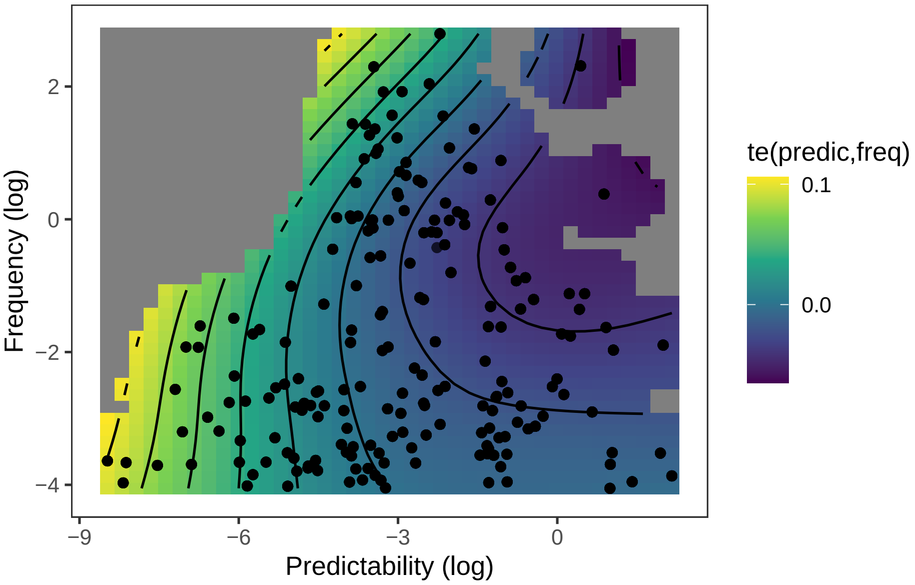

Figure 4: Plot of the interaction effect between predictability and frequency of the GAM model containing only the interaction between frequency and predictability (Equation 2). In the legend, te(predic,freq) refers to the predicted effect of the interaction effect. Thus, the brightness of the coloration denotes the strength of the interaction effect at the point in the graph. Brighter colors denote longer reaction times.

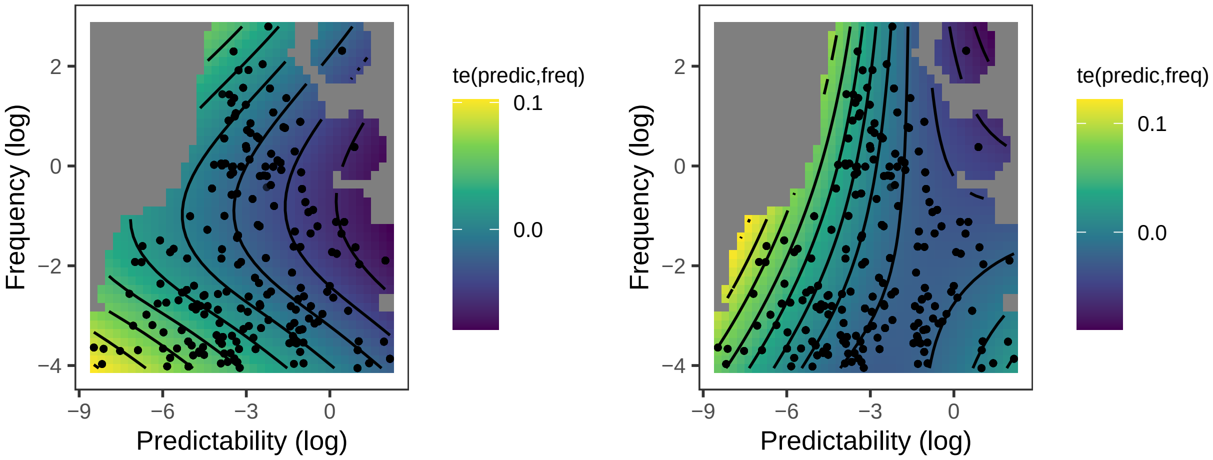

Figure 5: Plot of the interaction effect between predictability and frequency of the GAM model containing the interaction between frequency and predictability for phrasal vs nonphrasal verbs (Equation 3). Brighter colors denote longer reaction times. The left graph is the predicted effect for phrasal verbs (e.g., pick up), the right graph is the predicted effect for non-phrasal verbs (e.g., walk up).

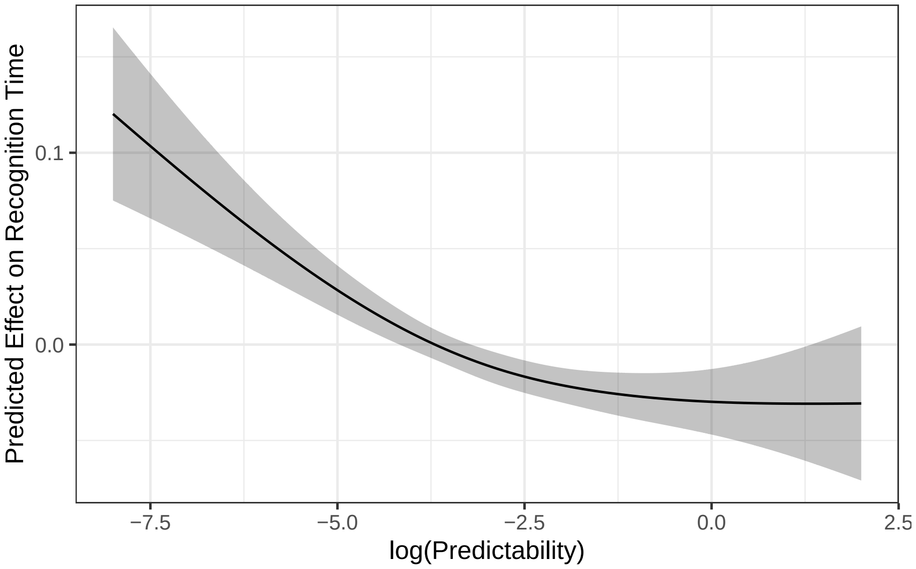

Figure 6: Plot of the predicted effect of log(predictability) on recongition time for the GAM model specified in Equation 4.

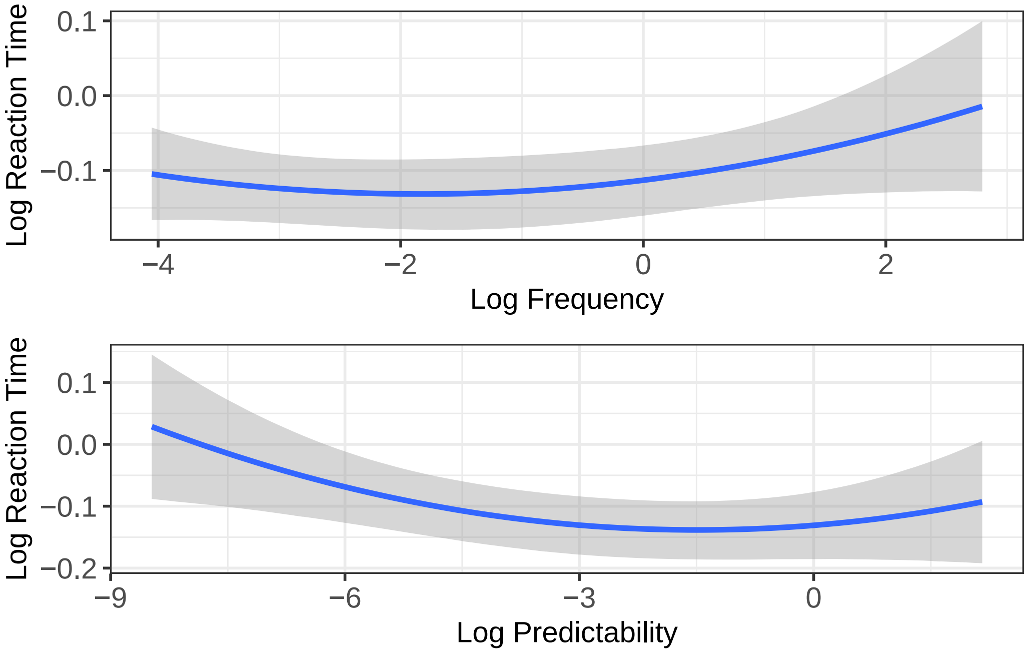

Figure 7: Visualization of the model results from Table 4 for frequency (top) and predictability (bottom). Frequencies are per million.

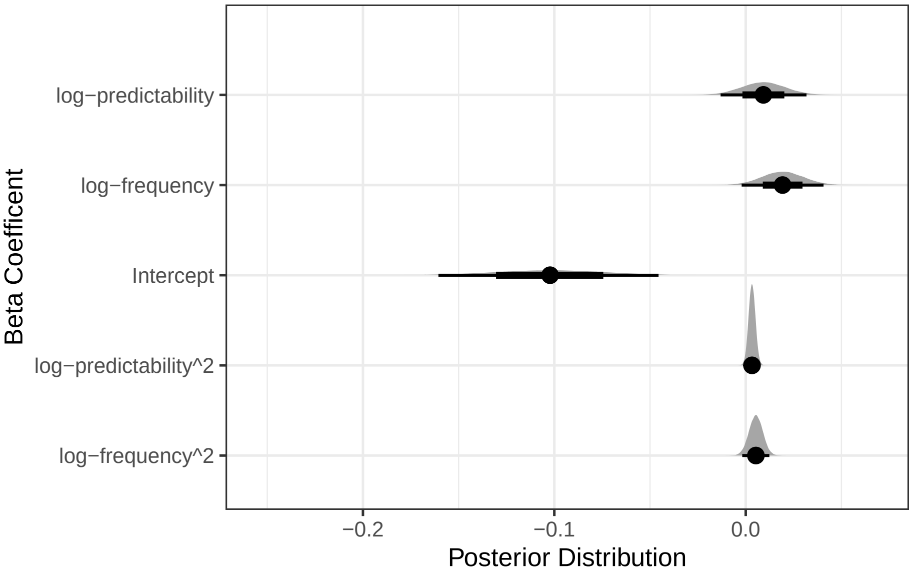

Figure 8: Plot of the posterior distribution for the beta value of each fixed-effect in the full Bayesian quadratic regression model (Equation 5). The y-axis contains the different fixed-effects and the x-axis contains the posterior distribution of beta values for the corresponding fixed-effect.

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,[7] as opposed to being two alternative realizations of a single construction. 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.

[7] 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.

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

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

Data Accessibility Statement

Data and analyses scripts can be found at the following url: https://anonymous.4open.science/r/Recognizability-Experiment-5B4B.

Acknowledgements

The authors would like to thank Dingyi (Penny) Pan for reading and offering comments on a draft of this paper. The authors would also like to thank Vsevolod Kapatsinski and Zara Harmon for providing helpful discussion about the nature of competition in lexical processing.

Competing interests

The authors have no competing interests to declare.

Authors' contributions

All authors were involved in conceptualization and methodology. The first and final authors were primarily involved in writing the original draft. All authors were involved in reviewing and editing the final draft.

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 [Publisher: Elsevier Inc.]. *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 [Series Title: Studies in Theoretical Psycholinguistics DOI: 10.1007/978-94-010-0355-1_3]. In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.). Springer Netherlands. http://link.springer.com/10.1007/978-94-010-0355-1_3

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations [Publisher: SAGE Publications Inc]. *Psychological Science*, *19*(3), 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal [PMID: 24403724 Publisher: Elsevier Inc.]. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28. https://www.jstatsoft.org/article/view/v080i01

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*, 1–26. 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

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use [PMID: 28503906]. *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 [PMID: 12757824]. *Trends in Cognitive Sciences*, *7*(5), 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Hampe, B. (2012). Transitive phrasal verbs in acquisition and use: A view from construction grammar [Number: 1]. *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 [PMID: 28830015 Publisher: Elsevier Inc.]. *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. [Publisher: American Psychological Association]. *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 [Publisher:

Cambridge University Press]. *Applied Psycholinguistics*, 1–17. https://doi.org/10.1017/S0142716423000486

Houghton, Z. N., & Morgan, E. (2023). Does predictability drive the holistic storage of compound nouns? [Issue: 45], *45*. https://escholarship.org/uc/item/7kz7w10b

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production [Publisher: Public Library of Science]. *PLOS ONE*, *7*(3), e33202. https://doi.org/10.1371/journal.pone.0033202

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

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

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

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

Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus, 169–174. https://aclanthology.org/P12-3029.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*, 522–533. https://www.academia.edu/download/68237640/Learning_Phonemes_Without_Minimal_Pairs20210721-21044-1t0fvya.pdf

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

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

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions [PMID: 27776281 Publisher: The Authors]. *Cognition*, *157*, 384–402. https://doi.org/10.1016/j.cognition.2016.09.011

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., Tenenbaum, J. B., & Goodman, N. D. (2009). Fragment grammars: Exploring computation and reuse in language [Accepted: 2009-03-31T05:00:03Z]. 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 [Publisher: Routledge _eprint: https://doi.org/10.1080/02643294.2019.1580189 PMID: 30806588]. *Cognitive Neuropsychology*, *36*(5-6), 216–219. https://doi.org/10.1080/02643294.2019.1580189

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior research methods*, *51*, 195–203.

Pinker, S. (1991). Rules of language. *Science*, *253*(5019), 530–535.

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

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

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

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

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

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

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. (2011). Seeing a phrase " time and again" matters: The role of phrasal frequency in the processing of multiword sequences [PMID: 21355667 ISBN: 0302006001]. *Journal of Experimental Psychology: Learning Memory and Cognition*, *37*(3), 776–784. https://doi.org/10.1037/a0022531

Starreveld, P. A., & La Heij, W. (1995). Semantic interference, orthographic facilitation, and their interaction in naming tasks. [Publisher: American Psychological Association]. *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

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*, 107–122.

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. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(1), 3–36.

Yi, B. W. (2002). Eumun hyeonsanggwa bindo hyogwa [the effect of usage frequency in phonology].

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 [ISBN: 9783319743943], 583–602. https://doi.org/10.1007/978-3-319-74394-3_20