Does Predictability Drive the Holistic Storage of Compound Nouns?

Zachary Nicholas Houghton (znhoughton@ucdavis.edu)
Department of Linguistics, University of California, Davis

Emily Morgan (eimorgan@ucdavis.edu)
Department of Linguistics, University of California, Davis

Abstract

Despite evidence that learners are storing a lot more than simple words, it is still unclear what determines whether a phrase is stored holistically. For example, storage could be driven by either phrasal frequency or by the mutual predictability of a phrase's component parts . Further, the processing consequences of holistic storage are also unclear. Given that sentence processing is incremental, how does recognition of individual words give rise to recognition of holistically stored phrases? The present study examines these questions. Specifically, participants are presented with sentences that contain compound nouns in locally plausible or locally implausible contexts. We examine whether participants are able to overcome local implausibility effects more easily if the compound nouns are highly predictable. We find that predictability does not overcome local implausibility effects, suggesting that either predictability is not driving holistic storage or that holistic storage driven by predictability does not facilitate comprehension in our task.

Keywords: Psycholinguistics; Sentence Processing

Introduction

Learning a language is not a trivial task. In order to be successful, learners must accurately segment the continuous speech stream into smaller segments, including phrases, words, morphemes, and phonemes. One of the main questions that arises out of this task is what exactly is the size of the units that learners are storing? That is, are they storing individual words, entire sentences, phrases, or some combination of all of these? One possibility is that learners store very little outside of words and idioms. For example, traditional theories have argued that learners don't store any more than they need to: they store only what they can't form compositionally using a set of rules, and generate everything else (e.g., Chomsky, 1975). For example, inflected words, such as *walked* would be generated by accessing the stored root, *walk*, and then applying a past tense rule that generates *walked* from the root. Similarly, a phrase like *I don't know* would be generated by accessing each of the individually stored words *I*, *don't*, and *know*.

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

It is also possible that what gets stored is somewhere in between these two extremes. For example, usage-based construction grammar approaches have posited that a lot more than just words are stored – including high frequency phrases – but rather than storing everything, or storing only the most basic units, that storage is driven by usage (Arnon and Snider, 2010; Bybee, 2003; Goldberg, 2003; Kapatsinski, 2018; Morgan and Levy, 2016; Tomasello, 2003). That is to say, the size of the units stored is driven by the statistical distribution of the language that the learner is producing and perceiving.

There is no shortage of evidence for the holistic storage of multi-word phrases. For example, high-frequency phrases, such as *I don't know*, have been shown to undergo phonetic reduction that isn't seen in other low or mid-frequency phrases containing *don't* (Bybee and Scheibman, 1999) suggesting that the representation of *I don't know* is separate from the representation of each of the individual words. In other words, the susceptibility of high-frequency phrases to phonological change is strong evidence that they may come to have a mental representation for the whole expression (i.e., holistic storage). This example is not an outlier either, there are many examples of high-frequency phrases undergoing phonetic reduction: *going to*, *want to*, *have to*, etc (Bybee, 2003).

Despite the clear evidence for the holistic storage of some multi-word units, however, it is still largely unclear what determines whether a unit is stored holistically. For example, it is possible that storage is driven by either **phrasal frequency** (co-occurrence rate, Bybee and Hopper, 2001) or by the mutual **predictability** of a phrase's component parts (i.e., how predictable the whole phrase is from part of the phrase, O'Donnell et al., 2011). For example, as previously stated, there is an abundance of evidence that high-frequency phrases are more susceptible to phonetic reduction than low-frequency phrases (Bybee, 2003; Bybee and Scheibman, 1999). Additionally, high-frequency phrases have been shown to lose the recognizability of their component parts relative to low-frequency phrases (Kapatsinski and Radicke, 2009). For example, *up* is harder to recognize in *pick up* than in *run up*.[1] On the other hand, in the learning literature in both humans and animals, there is significant evidence that learning is driven by prediction error as opposed to raw co-occurrence statistics. For example, Rescorla (1988) argued that learning in rats is not explained by simple co-occurrence statistics, but rather driven by error. That is, learning results from all of the cues in a given environment competing to predict the relevant outcome. In humans, Ramscar et al. (2013) demonstrated that in word learning, children rely on more than simple co-occurrence statistics but also on how *informative* – that is, how *predictive* – a cue is of an outcome (relative to other cues). Specifically, they demonstrated that children rely on not only co-occurrence rate, but also background rate (how often a cue is present without an outcome). In other words, assuming doors have a higher co-occurrence rate and lower background rate than all the other competing cues (e.g., brown, house, room) for the word *door*, then children will learn that doors are the best predictor of the word *door* (Ramscar et al., 2013).

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

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

1. **Novel Compound**

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

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

2. **Familiar Compound**

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

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

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

Thus there are two main problems that the present study aims to provide insight on: what exactly drives holistic storage, and what are the processing consequences of holistic storage? In Experiment 1, we first replicate Staub et al.'s (2007) experiment using a maze task (Boyce et al., 2020). In Experiment 2, we use the same methodology, but instead of using high (phrasal) frequency compound nouns, we use high *predictability* compound nouns (e.g., *peanut butter*). We ask whether the difference in reaction times between the locally implausible and plausible contexts differs depending on whether the compound noun is highly predictable or not.

Experiment 1

Methods

Participants

Participants were presented with sentences online via ibex farm (`github.com/addrummond/ibex`). 146 participants were recruited, however 30 participants were excluded for having an overall accuracy below 70%, leaving a total of 116 participants. All participants self-reported being native English speakers.

[1] Though for both of these contexts, predictability was not calculated, so it is unclear whether these were high or low predictability phrases.

Stimuli

Experiment 1 is a direct replication of Staub et al. (2007) using the maze task instead of eye-tracking[2]. The experimental sentences were sentences containing compound nouns from Staub et al. (2007) which varied upon two dimensions: local plausibility and familiarity. Locally plausible sentences were sentences in which the reading at the first noun was plausible and locally implausible sentences were sentences in which the reading at the first noun in the compound was implausible. Altogether, our stimuli consisted of 24 novel items, 24 familiar items (both from Staub et al.), and 188 filler sentences in order to avoid participants discerning the experimental design.

Procedure

Experiment 1 is a replication of Staub et al. (2007) using an A-Maze task (Boyce et al., 2020). In the A-maze task, participants are presented with the first word in the sentence and then have to correctly choose between an ungrammatical distractor word and the next word in the sentence. When participants select the correct word, they continue to the next word in the sentence until the sentence is finished. The distractor items were generated using the Gulordava language model (Gulordava et al., 2018).

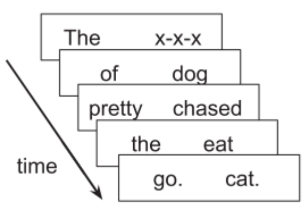


Figure 1: A visualization of the maze task, reproduced from Boyce et al. (2020)

Analysis

The data was analyzed using Bayesian linear regression models, as implemented in the *brms* package (Bürkner, 2018). We subsetted the data into two sets based on the region: one set for the first noun in the compound noun and one set for the second noun in the compound. For the purposes of this paper, we focus primarily on the N1 region. The primary dependent variable was log reaction time for both of these regions. The primary independent variables were plausibility and familiarity.

Results

The data was divided into two regions: the N1 region and the N2 region, which were the first and second noun in the compound noun respectively. The results of the Bayesian regression model for the N1 region are presented in Table 1 and in Figure 2a, and the results of the N2 region are presented in Table 2.

For the N1 region, there was an increase in reaction time for the implausible condition relative to the plausible condition. There was no such effect for familiarity. Additionally, there was no interaction effect between plausibility and familiarity.

At the N2 region, there was an increase in reaction time in the plausible condition and a decrease in reaction time in the familiar condition, but no interaction effect. However, plausibility did not mediate the effects of familiarity. That is to say, the size of the plausibility effect was not different for familiar versus novel compound nouns.

Discussion

Our results directly replicate Staub et al. (2007) using the Maze task, demonstrating the viability of this method for the tasks at hand. For the N1 region, while there was a clear increase in reaction time for items in the implausible condition, there was no interaction effect between plausibility and familiarity.

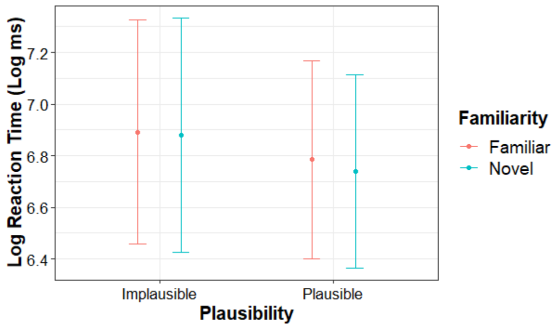
At the N2 region, the decrease in reaction time for familiarity is not surprising given that familiarity, as previously mentioned, was based on the frequency of the compound noun as a whole, however the increase in reaction time for the plausible condition is interesting, especially since the sentences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

It is possible that if *mountain lion* was stored holistically, then upon reading "Jenny heard the huge mountain...", the reader might have less difficulty with the local implausibility (relative to a low frequency compound noun) because they would predict *lion*, which eliminates the implausibility ("heard the mountain lion" is not implausible). However, we do not see this. Instead, the effect of plausibility is extremely similar for both familiar and novel items. One possible explanation for these results is that the familiar phrases are not necessarily stored. Instead storage might be driven by *predictability*. If this is the case then it would explain why we do not see this effect in Staub et al. (2007) or in Experiment 1, especially since all of the items used in Staub et al. (2007) are low predictability compound nouns (which we will operationalize in Experiment 2). We spend Experiment 2 exploring this question, using predictability instead of familiarity as the main independent variable.

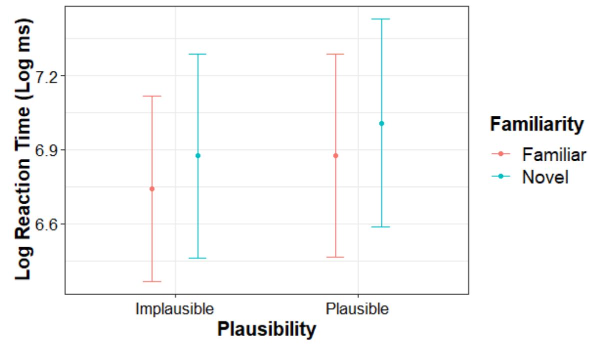
[2]The maze task was used due to the limitations of the COVID-19 pandemic

	Estimate	Est. Error	CI 2.5%	CI 97.5%
Intercept	6.82	0.02	6.78	6.87
Implausible	0.06	0.01	0.04	0.08
Familiar	0.01	0.02	-0.02	0.05
Implausible:Familiar	-0.00	0.01	-0.02	0.02

Table 1: Experiment 1: Regression analysis results for the N1 region.



(a) Experiment1: Plot of the N1 region. Lines indicate ±1 standard error.



(b) Experiment 1: Plot of the N2 region. Lines indicate ±1 standard error.

Experiment 2

Methods

Participants

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

Stimuli

We operationalized predictability through the odds ratio of the compound noun to the first word when that word is not followed by the second word in the compound noun which is exemplified in Equation 1.

$$\frac{\text{count}(\textit{peanut butter})}{\text{count}(\textit{peanut}) - \text{count}(\textit{peanut butter})} \tag{1}$$

In non-mathematical terms, Equation 1 quantifies how predictable the first noun is of the second noun (i.e., how likely the second noun is to follow after the first noun, relative to every other word that could follow). For example, the odds ratio of *peanut butter* would be the odds ratio of the compound noun – *peanut butter* – to the first noun – *peanut* – when *butter* does not follow it.

In order to collect the most predictable compound nouns, we searched the Google *n*-grams corpus (Michel et al., 2011) using the ZS Python package (Smith, 2014).[3] We then collected the compound nouns with the highest predictability values. We gathered a total of 37 compound nouns for our high predictability condition. We subsequently normed the sentences we created using the high predictability compounds, as well as the sentences from Staub et al. (2007) which we confirmed were all low predictability compounds relative to our compound nouns. We provided participants with each item in four conditions (plausible sentence, through the first noun; plausible sentence, through the second noun in the compound; implausible condition, through the first noun in the compound; implausible condition through the second noun in the compound) and asked participants to rate each sentence in terms of how well the last word fit in the sentence. No participant rated more than one version of each sentence. The mean values for each condition are as follows: plausible, through the first noun: 5.58 (sd = 0.78); plausible, through the second noun: 5.41 (sd = 0.71); implausible, through the first noun: 3.13 (0.63); implausible, through the second noun: 5.47 (sd = 0.82).

After norming, we selected sentences such that the difference in plausibility values between the plausible and implausible conditions were roughly the same for the high predictability and low predictability conditions. After accounting for this, we ended up with 21 high predictability and 21 low predictability items (which were taken from Staub et al., 2007), for a total of 42 items. Lastly, in order to avoid participants discerning the experimental design we also included 188 filler items.

Procedure

Our procedure was the same as in experiment 1.

[3]The code we used to search the bigrams corpus along with the materials and results are all freely available via: https://github.com/ucdavis/predictability

	Estimate	Est. Error	CI 2.5%	CI 97.5%
Intercept	6.87	0.02	6.82	6.92
Implausible	-0.07	0.02	-0.11	-0.05
Familiar	-0.07	0.02	-0.11	-0.03
Implausible:Familiar	0.00	0.01	-0.01	0.02

Table 2: Experiment 2: Regression analysis results for the N2 region.

Analysis

The data was analyzed using Bayesian linear mixed-effects models, as implemented in the *brms* package (Bürkner, 2018). The primary dependent variable was log reaction time for both of these regions (following Boyce et al., 2020). The primary independent variables were plausibility and predictability. Reaction time was modeled as a function of plausibility and predictability, along with their interaction, with maximal random effects (following Barr et al., 2013).

Results

The results of the Bayesian regression models for the N1 region are presented in Table 3 and Figure 3, and the results of the N2 region are presented in Table 4.

With regards to the N1 region, Table 3 presents the results of the analysis we ran with predictability as a continuous predictor (operationalized as the log odds ratio). Our results demonstrate that, similar to experiment 1, there was an increase in reaction time for the implausible condition, but no effect for predictability or the interaction between the two.

With regards to the N2 region, Table 4 presents the results of the analysis we ran with predictability as a continuous predictor (operationalized as the log odds ratio). Our results, as in Experiment 1, demonstrate an increase in reaction time in the plausible condition and and a decrease in reaction time in the high-predictability condition, but no interaction effect between plausibility and predictability.

Discussion

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

General Discussion and Conclusion

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

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

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

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

Another possibility is that the high-predictability compound nouns are stored holistically, but the present study did not succeed in eliciting an effect for them due to the choice of the task. For example, in the Maze task participants are forced to make a more active choice than in a naturalistic reading paradigm (i.e., they have to actively select the correct word). It is possible that this choice forces participants into a specific commitment of a sentence. Additionally, unlike eye-tracking, in the Maze

	Estimate	Est. Error	CI 2.5%	CI 97.5%
Intercept	6.88	0.03	6.82	6.93
Implausible	0.07	0.01	.04	0.10
LogOddsRatio	0.01	0.01	-0.01	0.02
Implausible:LogOdds	0.00	0.00	-0.01	0.01

Table 3: Experiment 2: regression analysis results for the N1 region with predictability as a continuous predictor (log odds ratio).

	Estimate	Est. Error	CI 2.5%	CI 97.5%
Intercept	6.80	0.03	6.75	6.85
Implausible	-0.07	0.01	-0.10	-0.05
LogOddsRatio	-0.03	0.01	-0.05	-0.02
Implausible:LogOdds	0.00	0.00	-0.01	0.01

Table 4: Experiment 2: regression analysis results for the N2 region with predictability as a continuous predictor (log odds ratio).

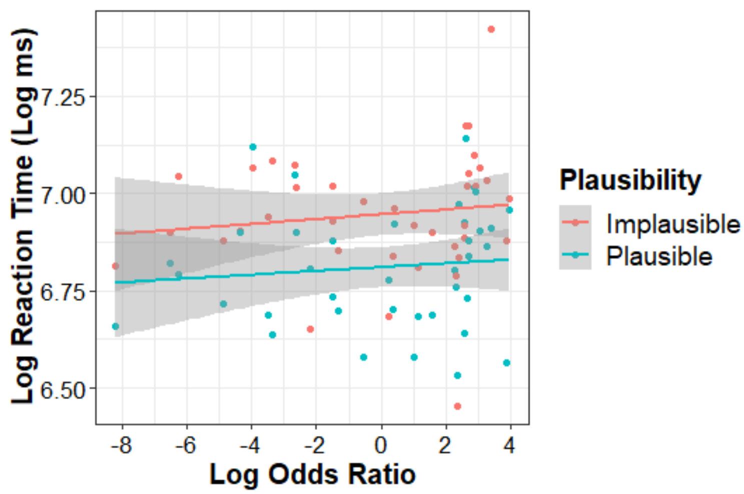


Figure 3: Plot of the N1 region with predictability as a continuous variable, shading indicates ±1 standard error. Our linear regression model suggested that there was a significant decrease in reaction time for both the implausible condition and the high-predictability condition, but no interaction effect between the two. Notice that the lines are nearly parallel. If there was an interaction effect we would see a notable lack of parallelism between the two lines.

task participants are comparing two interpretations of a sentence (i.e., they're comparing each word on the screen as a potential continuation of the sentence), which may have unintended effects that don't reflect the processes of sentence processing in a more naturalistic environment. When COVID permits, we look forward to replicating the present study using eye-tracking in order to assess whether the results are task-specific.

One last possibility is that, like the previous possibility, the high-predictability compounds are stored holistically, but there is no facilitatory effect in the processing of the first noun in the compound noun. This would certainly beg the question, however, of what the processing consequences of holistic storage are. Perhaps the primary advantages to holistic storage are in production rather than processing. For example, it is possible that it is quicker to access and produce a stored compound noun, but during language comprehension the processing system may avoid accessing words that we haven't heard yet since it may be riskier (for more debate ove the role of prediction in comprehension see Ferreira and Chantavarin, 2018; Onnis and Huettig, 2021).

In summary, the present study contributes to the current theories of sentence processing by demonstrating that predictability may not be the driving factor behind holistic storage , however given the lack of research demonstrating the specific processing consequences of holistic storage, it is possible that rather than predictability not driving holistic storage, either our task doesn't elicit a measurable effect of holistic storage, or holistic storage of a compound noun just doesn't facilitate the processing of the first noun in compound nouns. Either way, a follow-up eye-tracking study would shed a great deal of light on these questions by demonstrating the generalizability (or perhaps a lack thereof) of our current findings.

References

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

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

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

Boyce, V., Futrell, R., and Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

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

Bybee, J. and Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in english. 37(4):575–596.

Bybee, J. L. and Hopper, P. J., editors (2001). *Frequency and the Emergence of Linguistic Structure*. John Benjamins.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

Chomsky, N. (1975). Reflections on language.

Ferreira, F. and Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current directions in psychological science*, 27(6):443–448.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically.

Kapatsinski, V. (2018). *Changing Minds Changing Tools: From Learning Theory to Language Acquisition to Language Change*. The MIT Press.

Kapatsinski, V. and Radicke, J. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. *Formulaic language*, 2:499–520.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Morgan, E. and Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, 157:384–402.

O'Donnell, T., Snedeker, J., Tenenbaum, J., and Goodman, N. (2011). Productivity and reuse in language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Onnis, L. and Huettig, F. (2021). Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled'from memory or compositionally constructed on the fly? *Brain Research*, 1772:147674.

Ramscar, M., Dye, M., and Klein, J. (2013). Children value informativity over logic in word learning. *Psychological science*, 24(6):1017–1023.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American psychologist*, 43(3):151.

Smith, N. J. (2014). Zs: A file format for efficiently distributing, using, and archiving record-oriented data sets of any size. *Manuscript submitted for publication. School of Informatics, University of Edinburgh. Retrieved from http://vorpus. org/papers/draft/zs-paper. pdf.*

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

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.