AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review


ISSN: 0033-295X

2021, Vol. 128, No. 6, 1112–1144
https://doi.org/10.1037/rev0000305

# A Theory of Repetition and Retrieval in Language Production

Zara Harmon[1, 2] and Vsevolod Kapatsinski[3]
[1] Institute for Advanced Computer Studies (UMIACS), University of Maryland
[2] Department of Linguistics, University of Maryland
[3] Department of Linguistics, University of Oregon

Repetition appears to be part of error correction and action preparation in all domains that involve producing an action sequence. The present work contends that the ubiquity of repetition is due to its role in resolving a problem inherent to planning and retrieval of action sequences*: the Problem of Retrieval*. Repetitions occur when the production to perform next is not activated enough to be executed. Repetitions are helpful in this situation because the repeated action sequence activates the likely continuation. We model a corpus of natural speech using a recurrent network, with words as units of production. We show that repeated material makes upcoming words more predictable, especially when more than one word is repeated. Speakers are argued to produce multiword repetitions by using backward associations to reactivate recently produced words. The existence of multiword repetitions means that speakers must decide where to reinitiate execution from. We show that production restarts from words that have seldom occurred in a predictive preceding-word context and have often occurred utterance-initially. These results are explained by competition between preceding-context and top-down cues over the course of language learning. The proposed theory improves on structural accounts of repetition disfluencies, and integrates repetition disfluencies in language production with repetitions observed in other domains of skilled action.

*Keywords:* disfluency, surprisal, cue competition, sequential behavior, skilled action



When we type *repeating yourself is* into Google, the likely continuations indicate that repetition is thought to be a sign of disorder and disease, anything from dementia and Alzheimer's to obsessive-compulsive disorder and even alcoholism. Nonetheless, everyday action involves extensive repetition. For example, when experiencing a speech error, one invariably repeats the entire word containing the error, and often more than that; for example, *I like to listen to the newsp− . . . the radio in the morning*. Even if the error is localized to the end of the word, as in *hypothethith*, the speaker always restarts the word from the beginning to correct it. Repetition of pre-error material is also part of error correction outside of language production. For example, a competitive dancer who makes a mistake in their routine is likely to restart the routine from the beginning rather than from the location of the error. Repetition occurs in the absence of error as well, as the agent is trying to plan or retrieve the next action. Consider an athlete trying to complete an action that requires significant planning and retrieval time, such as a basketball free throw, a golf putt, or a tennis serve. In all such cases, some repetitive behavior—a pre-performance routine—is usually performed prior to each attempt (see Dömötör et al., 2016, for a summary). Pre-performance routines are especially likely, and particularly prolonged, when the attempted action is difficult. Similarly, in speech, repetitions are common before difficult-to-access words (Harmon & Kapatsinski, 2015). These types of repetitions are usually called *repetition disfluencies*, illustrated in (1)–(3). In these examples, the brackets isolate the repetition and the plus sign marks the interruption point. The words preceding the interruption point are indexed with negative numbers indicating distance from the interruption point, while the word following the interruption point is indexed with 1.

1. But it's really$_{-3}$ $a_{-2}$ [big$_{-1}$, + big,] decision$_1$ as to, you know, when to do it.
2. It was just $a_{-3}$ [change$_{-2}$ of$_{-1}$, + change of] location$_1$.
3. I'm doing basically system design work and, uh, implementation work [for$_{-3}$ the$_{-2}$ speech$_{-1}$, + for the speech] group$_1$.

The present article proposes that repetition of this kind is functional: rather than being mere perseveration error, it helps solve what we term the *Problem of Retrieval*. That is, repetition helps access upcoming action(s) by reactivating one or more preceding words. The proposed theory answers several questions about repetition

This article was published Online First July 5, 2021.

Zara Harmon https://orcid.org/0000-0001-9692-9380

We are grateful to Gary Dell, Gordon Logan, and an anonymous reviewer for their insightful comments and most helpful feedback on the article. We are indebted to Michal Young for help with Python coding. We also thank Naomi Feldman, Nazbanou Nozari, Eric Pederson, Tyler Kendall, and Colin Bannard, as well as members of the Computational Linguistics and Information Processing (CLIP) Lab at the University of Maryland and the Cognitive Linguistics Workgroup at the University of Oregon for helpful comments and discussion of the ideas at different stages of the article.

The work has benefited from presentation at a variety of venues including the International Cognitive Linguistics Conference, 2015, Conceptual Structure, Discourse, and Language, 2014, the Acoustical Society of America, 2016, Annual Meeting of the Psychonomic Society, 2018, and Cognitive Science Society, 2020. We thank the audiences at these conferences for helpful discussions. Some of the ideas and data discussed here have been published in Harmon and Kapatsinski (2015, 2016, 2020) and Kapatsinski (2018b).

Correspondence concerning this article should be addressed to Zara Harmon, Department of Linguistics, University of Maryland, 1401 Marie Mount Hall, College Park, MD 20742, Unites States. Email: zharmon@umd.edu

disfluencies: why do they occur, why they sometimes involve repeating more than one word, as in (2)–(3), and finally, what determines how many words are repeated: why does the speaker in (1) reinitiate production from word $w_{-1}$ while the one in (3) reinitiates it from word $w_{-3}$? We suggest that a multiword repetition is even more helpful than a single-word repetition for retrieving the upcoming word. However, a multiword repetition requires reactivation of recently produced words to execute. We implement these retrieval and reactivation processes in a computational model that captures the effects of linguistic experience on a word's accessibility in context.

The remainder of this article is organized as follows. We first present an overview of the modeling approach. We then present the three hypotheses that constitute our theory. The *Facilitation Hypothesis* maintains that repetitions facilitate accessing the upcoming item. This hypothesis relies on the assumption that words are activated by predictive preceding contexts. The *Reactivation Hypothesis* specifies the process by which words preceding the interruption point are reactivated and repeated. It claims that previously produced words must be reactivated to be re-produced, and that this reactivation process uses the words that follow as cues. The *Initiation Hypothesis* addresses the mechanism responsible for the speaker's choice of word from which to restart speech. It claims that, after an interruption, speech is restarted from words that have occurred relatively unexpectedly in the speaker's prior experience. The Initiation Hypothesis attributes this effect to *Cue Competition* between preceding context and top-down cues: initiation from words that tend to occur in predictive preceding-word contexts is relatively unlikely because such words have a weaker association with top-down cues. We then show how the three hypotheses work together to account for repetition behavior above and beyond previously proposed explanations. We conclude by discussing how the proposed theory can explain crosslinguistic variability in repetition behavior and outlining a number of novel predictions that could be tested in other domains of skilled action.

## Modeling Approach

We trained a recurrent neural network (van Schijndel & Linzen, 2018) on Switchboard (Godfrey et al., 1992), a 1.7-million-word corpus of American English conversations that is the source of our disfluency data. Here, we use the long short-term memory (LSTM) variant of the recurrent network architecture (Hochreiter & Schmidhuber, 1997) with continuous updating of weights, as implemented in pytorch (van Schijndel & Linzen, 2018).

Recurrent neural networks embody the hypothesis that words are activated by preceding contexts that predict their occurrence. For this reason, recurrent networks have been widely used as models of language processing (Chang et al., 2006; Dell et al., 1993; Elman, 1990) and action sequencing (Botvinick & Plaut, 2004, 2006; Cooper et al., 2014). They produce state-of-the-art results in capturing between-word dependencies in sentence processing (e.g., Gulordava et al., 2018; Linzen et al., 2016; van Schijndel & Linzen, 2018) and the magnitude of associative priming between co-occurring words (Moss et al., 1994). In work on disfluencies, Dammalapati et al. (2019) have recently shown that recurrent networks appear to capture the severity of the Problem of Retrieval a word poses in context, by predicting how likely a disfluency is to occur before a word.

We take these results to suggest that recurrent networks can capture the strength with which a word is activated by a particular context. In such a network, a word is activated by a context to the extent that the context predicts the word's occurrence. This assumption is central to the proposed theory of repetition and retrieval. The Facilitation Hypothesis assumes that words are activated by predictive preceding contexts. The Reactivation Hypothesis proposes that words are reactivated using following contexts. The Initiation Hypothesis proposes that speakers tend not to restart production from words that occur in predictive preceding-word contexts. Therefore, testing the three hypotheses that constitute the proposed theory relies on quantifying the degree to which words are activated by a particular context.

## The Facilitation Hypothesis: Repetitions Help Solve the Problem of Retrieval

At least since Lounsbury (1954) and Goldman-Eisler (1957, 1958, 1968), disfluencies have been argued to buy the speaker planning time. That is, disfluencies occur when the speaker is having trouble deciding what to say next. The most common source of difficulty appears to be lexical retrieval (Hieke, 1981). Thus, disfluencies tend to occur before the kinds of words that are especially difficult to retrieve: Words that are infrequent or low-probability in the current context (Beattie & Butterworth, 1979; Dammalapati et al., 2019; Goldman-Eisler, 1957, 1958, 1968), and words that have many semantic competitors (Harmon & Kapatsinski, 2015; Hartsuiker & Notebaert, 2010; Schachter et al., 1991; Schnadt, 2009). This pattern is so robust in adult speech that listeners learn to use it to guide comprehension, starting as early as age 2 (Kidd et al., 2011).

Branigan et al. (1999) suggest that repetitions are particularly likely to be triggered by the need to buy time while remaining committed to the current speech plan, whereas other disfluencies are more likely to be triggered by the need to alter the current plan. Thus, repetition appears to occur when the speaker is having trouble retrieving the next word from their mental lexicon but believes that the word they are trying to retrieve is appropriate to the current context. In other words, the speaker is facing something resembling a tip-of-the-tongue state (as defined by Brown, 1991): The correct word is partially activated by the production plan but not strongly enough to be selected for execution. This is the Problem of Retrieval. In this article, we argue that repetitions solve the Problem of Retrieval, resolving the tip-of-the-tongue state. We call this proposal the Facilitation Hypothesis.

According to the Facilitation Hypothesis, repetition helps the intended continuation win the race by cueing words that tend to follow the one(s) that the speaker has just said. This hypothesis relies on three assumptions regarding lexical processing: (a) co-occurring words are associated with each other, (b) repeating a word provides it with additional activation, and (c) activating a word results in activation of its associates (e.g., Moss et al., 1994). These assumptions appear relatively uncontroversial. We know that repetition boosts activation of the repeated word because speakers show both identity and homophone priming in production. In particular, producing a form makes it more likely that it would be produced again in the future (Barry et al., 2001; Burke et al., 2004; Ferreira & Griffin, 2003), and wordforms are articulated more quickly when they have been recently produced (Shields & Balota, 1991; Turnbull, 2019). We know that predictive preceding contexts activate the associated words because predictable words are easier to access, resulting in greater fluency (Goldman-Eisler, 1957, 1958) and shorter articulation duration (Turnbull, 2019).

Given these three assumptions, repetition of a word should provide associated words with additional activation, making them easier to

access. Since speakers do not usually produce sequences of unrelated, unassociated words (or else they would become associated), repetition should then be generally helpful for solving the Problem of Retrieval. However, because repetitions tend to precede words of low predictability given the preceding context (Dammalapati et al., 2019; Goldman-Eisler, 1957, 1958), one may worry that repeating the preceding context would be unhelpful specifically where it tends to occur. That is, the repeated preceding context might cue highly predictable continuations, exactly those which are unlikely to be intended by the speaker while facing a retrieval problem. If so, repeating this context would make the intended upcoming words less accessible, contradicting the Facilitation Hypothesis. In this section, we test this possibility by investigating whether the repeated context for an upcoming word, $w_1$, helps the network predict its occurrence.

The Facilitation Hypothesis makes three predictions about behavior of the LSTM network. First, the network should predict $w_1$ better if it has access to the identity of the preceding word, $w_{-1}$. If this is true, then repeating even one word would be helpful to retrieve the partially activated $w_1$ relative to other words. Second a longer context should help the network predict $w_1$. If this is true, then repeating more words would generally be more helpful for solving the Problem of Retrieval. We expect this to be the case even for the relatively unpredictable words that follow disfluencies. This greater predictiveness of longer contexts is one motivation for the existence of multiword repetitions, providing an explanation for why the speaker would go to the trouble of re-producing the past. Third, we test whether there is some degree of alignment between how many words are repeated and how many words would be most helpful to repeat. The existence of such an alignment would suggest that whatever mechanisms lead the speaker to repeat a certain number of words result—at least in English—in repetitions that are particularly helpful for solving the Problem of Retrieval. We return to the question of whether this alignment plays a causal role in determining the number of words repeated later in the article, where we pit it against other influences on the process in a regression model.

## Method

Our simulations used the specific LSTM architecture and parameter settings previously used by van Schijndel and Linzen (2018) to model sentence processing in reading. The model produces an output activation for every word in the corpus given the preceding context. Output activations were transformed into choice probabilities of words given contexts using the softmax function. Probabilities were then transformed into surprisal values (Information content, denoted by $I$), where $\mathrm{I}(w|\text{context}) = -\log(p(w|\text{context}))$. Surprisal is the amount of information (in bits) about the identity of a word gained by observing the word in context (Chater, 1996; Levy, 2008), that is, how unexpected the word's occurrence is to the network after the particular preceding context. We predict that adding words to the preceding context available to predict $w_1$ would reduce the network's surprisal at $w_1$. We refer to the difference in surprisal at $w_1$ given a shorter context minus a longer context as relative surprisal or gain in predictability from retrieving the word that distinguishes the two contexts. For example, relative surprisal of $w_{-2}w_{-1}$ vs. $w_{-1}$ is defined as $I(w_1|w_{-1}) - I(w_1|w_{-2}w_{-1})$. Greater relative surprisal means that the longer context is much more predictive than the corresponding shorter context.

As in van Schijndel and Linzen (2018), we used two LSTM layers with 200 hidden units in each, a learning rate multiplier for the gradient equal to 20, a cross-entropy loss function, 40 as the maximum number of training epochs with training stopping if loss remains constant for three consecutive epochs, 20% dropout during training, and 35 words as the context length to backpropagate error. We ran 10 models initialized with different random seeds and different order of corpus sentences in training. Correlations between the surprisal weights of these models given a context were high (mean $r = .94$; range: .88–.95), showing that the results are robust to these sources of random variance. We selected the model with the default random seed and the original order of corpus sentences, which had an average mean correlation to the others (mean $r = .94$).

The corpus was fed to the model word by word, with utterance boundaries marked by a special start symbol. The model generated a surprisal value for each word it encountered. The repetition disfluency contexts were extracted from Switchboard and were used to test the model's ability to predict the words that follow repetition disfluencies. In testing the model, we manipulated the length of the context available to the model to predict each word, that is, whether the model was presented with no preceding words, one preceding word, two preceding words, or three preceding words. If repeated units are predictive of upcoming words, we expect surprisal for words following a disfluency to decrease with increasing context length. Also, of interest is whether the predictiveness of a two-word or three-word context relative to a shorter context correlates with how many words the speaker actually repeats. This allowed us to observe the correlation between the number of words repeated by the speaker and how much a repetition of that length would reduce surprisal compared to a shorter repetition.

We considered two measures of surprisal: Simple surprisal of the word that follows a disfluency ($w_1$) given the preceding context, and the difference in surprisal between $w_1$ and its nearest semantic competitor given context. We defined nearest competitor using Latent Semantic Analysis (LSA; Landauer & Dumais, 1997; as implemented in the lsa package in R; Wild, 2015), with "documents" defined as conversations from the corpus. LSA performs a principal components analysis of the word-by-document matrix, and then uses the resulting principal components as dimensions of a similarity space. Words are similar if they tend to occur in the same documents or if they tend to co-occur with the same words. The crucial property of LSA for the present study is that it is a *bag-of-words* model that does not take into account word order or distance between words in a document. It therefore forms a valid baseline to investigate the importance of adding associative contextual cues. More recent models add information about the local context to semantic representations, making it difficult to identify an independent contribution of contextual information (e.g., Aina et al., 2019; Boleda, 2020). We also coded whether the nearest semantic competitor matched $w_1$ in syntactic category.[1]

### Database

One-, two-, and three-word repetitions were retrieved from the Switchboard Corpus using Python regular expressions. The same

[1] LSA semantic similarities often don't match human intuitions. For this reason, we performed supplementary analyses in which we coded whether the nearest semantic competitor assigned by the LSA appeared to be strongly related to $w_1$ in meaning, and removed cases where this was not the case. The coding was blind to how many words were repeated, and was performed prior to analyzing the data. As reported in the next footnote, the results were virtually identical to the results reported in the text for semantic competitors matching $w_1$'s syntactic category.

**Table 1**
*Mean(SD) Surprisal Values From the LSTM as a Function of the Number of Words Available to Predict $w_1$ and the Number of Words Repeated*

| Number of Words Repeated | Zero available | One available | Two available | Three available |
|---|---|---|---|---|
| One repeated | 47.37(4.85) | **10.95(4.90)** | 9.17(4.39) | 8.59(4.20) |
| Two repeated | 47.54(4.90) | 11.81(4.83) | **8.73(4.27)** | 8.22(4.11) |
| Three repeated | 47.38(4.89) | 11.62(4.72) | 8.95(4.31) | **7.61(4.14)** |
| Overall | 47.42(4.87) | 11.22(4.88) | 9.04(4.35) | 8.43(4.18) |

*Note*. The columns indicate how many words were available to the model to predict $w_1$. The rows indicate how many words the speaker repeated. Bolded values are surprisal values obtained when the number of words available to the model to predict $w_1$ matches with the number of words repeated by the speaker. LSTM = long short-term memory. $w_1$ = the word that follows the disfluency.



database of disfluencies was used for all analyses in the present article. Therefore, disfluencies needed to meet a relatively stringent set of criteria to be included. We excluded one-word repetitions that started within two words of the preceding clause boundary and two-word repetitions that started within one word of the preceding clause boundary. Repetition disfluencies do not span clause boundaries, so this exclusion ensured that the speaker always had the option to produce a longer repetition than the one they actually produced, and that more than one word was available to predict the upcoming word. Complex disfluencies, that is, cases in which other restarts and repetitions immediately preceded or followed the repetition disfluency, as well as abandonments, were also excluded. Finally, we removed instances where the preceding context contained two-word sequences tagged in the corpus as discourse markers such as *I think*, *I guess*, and *you know* because these may function as single words. Speech was never restarted from inside such a unit, just as there are no restarts from word-internal locations. An additional reason to exclude discourse markers is that they may themselves function to buy time for planning upcoming speech, which makes the resulting utterance a complex disfluency. Note that our results would be stronger if they were included in the dataset. Thus, their exclusion is a conservative choice given our hypotheses. We also removed utterances for which we did not have duration information. The final sample included we arrived at 2,988 one-word repetitions, 1,160 two-word repetitions, and 294 three-word repetitions.

### *Statistical Analysis*

We evaluate whether, on a typical occasion of struggling to retrieve an upcoming word, a repetition is likely to be helpful. The results in this section were analyzed using simple *t*-tests rather than a mixed-effects regression model, that is, they do not include a random effect of word. While observations coming from the same word are not independent statistically, we believe that the speaker treats them as independent in learning whether repetitions are helpful. That is, even if all of the speaker's experience involved a single $w_1$, but repetitions were helpful to access it, the speaker would still learn to employ repetitions to help access their one word.

## Results

In the first set of analyses, we examined whether surprisal of $w_1$ in context differed based on the length of the context available to the model to predict the word. Table 1 reports the surprisal values for $w_1$ depending on whether the speaker repeats one, two, or three words. The results provide support for the Facilitation Hypothesis. First, there is a large (36 bit) decrease in surprisal (i.e., an increase in predictability) from having even the single preceding word available to predict $w_1$ (zero available vs. one available; $t(4441) = 813.31$, $p < .0001$). Second, the model can predict upcoming words better given a longer context, indicating that longer repetitions tend to be more helpful than shorter ones (one vs. two: $t(4441) = 59.68$, $p < .0001$; two vs. three: $t(4441) = 27.93$, $p < .0001$), although adding words to the context yields diminishing returns: $w_{-2}$ is worth 2.2 bits of information, while $w_{-3}$ is only worth 0.6 bits.

There is some degree of alignment between how many words are repeated and how many words would be helpful to repeat. Even though repeating more words is generally helpful for retrieving the future, a context of a certain length is more predictive when it is repeated than when it is not (within each column in Table 1, the bold values are on average 0.88 bits lower than plain text ones; $t(17168) = 10.12$, $p < .0001$). Figure 1 illustrates this point: repeated words tend to decrease surprisal about the future more than words in the same position that are not repeated.

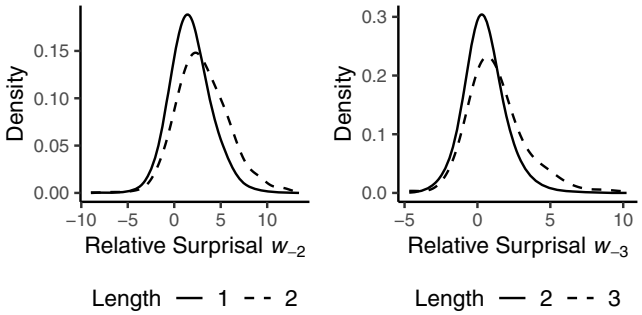
In the preceding analysis, the activation of the intended continuation was measured relative to all possible continuations. One may therefore argue that the preceding context helps predict $w_1$ in that analysis only because it is ruling out continuations that are unlikely to compete with the actual continuation for selection, i.e., words that would be completely inappropriate in the context. To rule out this possibility, we investigated whether the preceding context improves predictability of a word relative to its closest semantic competitor, as defined by LSA.

The difference in surprisal of the word and its nearest semantic competitor was 4 bits out of context (in favor of the observed continuation). Adding one word of preceding context ($w_{-1}$) increased this difference by 2.3 bits (±0.1), while adding $w_{-2}$ increased it by 1.9 bits (±0.1), and $w_{-3}$ increased it by an additional 0.64 bits (±0.06), all differences significant at $p < .0001$ [$t(4087) = 40.76$, 44.88, and 21.77 respectively].[2] These results on the difference in surprisal between $w_1$ and its closest semantic competitor, largely align with the results on

[2] The results were very similar when the competitor was required to match the syntactic category of the target word, indicating that the preceding-word context contributes to word choice beyond identifying its syntactic category. Adding one word of context ($w_{-1}$) increased the difference in surprisal between target and competitor from 1.5 to 3 bits (by 1.5 bits), adding $w_{-2}$ increased it by an additional 1.5 bits, while $w_{-3}$ increased it by 0.61 bits; all differences significant at $p < .0001$ [$t(1104) = 14.42$, 18.03, and 10.82 respectively]. Even when comparisons are restricted to words that both match in syntactic category and are intuitively highly related semantically, the results remain qualitatively unchanged: $w_{-1}$ increases the advantage of the target over the competitor by 1.19 bits (from 1.29), $w_{-2}$ adds 1.36 bits, and $w_{-3}$ adds 0.51 bits. All differences are significant at $p < .0001$ [$t(828) = 10.21$, 14.53, 8.21 respectively].

**Figure 1**

*The Distribution of Relative Surprisal or Gain in Predictability of $w_1$ From Adding an Additional Word to the Cueing Context as a Function of how many Words are Repeated (Length): Repeated (Dashed Lines) vs. Not Repeated (Solid Lines)*




*Note.* Higher levels of relative surprisal / gain in predictability mean that the addition of a word to the context strongly increases predictability (decreases surprisal) of the future ($w_1$). Left panel: gain in predictability from adding $w_{-2}$: $I(w_1|w_{-1}) - I(w_1|w_{-2}w_{-1})$. Right panel: gain from adding $w_{-3}$: $I(w_1|w_{-2}w_{-1}) - I(w_1|w_{-3}w_{-2}w_{-1})$. An additional word of context helps predict the future more when it is repeated than when it is not.

*simple surprisal* of $w_1$ reported above (bottom row of Table 1). Knowing the preceding word helps predict the upcoming word, and knowing a longer context helps even more.[3] There are also diminishing returns from knowing $w_{-3}$ compared to knowing $w_{-2}$ or $w_{-1}$. However, the decrease in surprisal from knowing $w_{-1}$ is more modest when measured relative to the nearest competitor than when measured relative to all words. This shows that much of the reduction in absolute surprisal of $w_1$ resulting from knowing $w_{-1}$ is shared with $w_1$'s likely semantic competitors. As a result, in discriminating close semantic competitors, augmenting the context with $w_{-2}$ improves predictability almost as much as augmenting it with $w_{-1}$. Therefore, a multiword repetition can be substantially more helpful than a one-word repetition for accessing the future. The results discussed above indicate that longer contexts are generally more predictive of the identity of $w_1$ than shorter contexts. At the same time, the length of the context does not strongly influence *which* $w_1$ tokens are more predictable: surprisal scores given a one-word context strongly correlate with surprisal scores given a two-word context or a three-word context (Table 2; $N = 4{,}442$). All surprisal scores from the LSTM also strongly correlate with surprisal in a simple bigram model in which each word predicts the word that immediately follows—that is, the inverse of log transitional probability. These correlations are almost as high as intercorrelations between LSTM models initialized with different random seeds or presented with the corpus sentences in a different order ($.89 < r < .94$), indicating that LSTM surprisal largely reduces to transitional probability (the probability of a word given the preceding word), which captures more than 90% of the variance shared between the LSTM models. Thus, predictiveness of a context, to a very large extent, tracks predictiveness of its final word. This means that the effects of contextual predictability on number of words repeated, presented below, are largely robust to assumptions about how context is represented, and how much of the context is used to predict the future. To keep LSTM surprisal distinct from surprisal in the bigram model, we will refer to the latter as log transitional probability.

**Table 2**

*Correlations (Pearson r) Between Surprisal Values*

| Context Length | One word | Two words | Three words |
|---|---|---|---|
| One word | — | — | — |
| Two words | .87 | — | — |
| Three words | .81 | .94 | — |
| $I(w_1\|w_{-1})$ | .85 | .87 | .85 |

*Note.* The top three rows show correlations between LSTM surprisal values depending on the length of context available to the model to predict $w_1$ (one, two or three words). The bottom row shows correlations of LSTM surprisal with surprisal in a simple bigram model. LSTM = long short-term memory.

## Discussion

These results indicate that repeated words are predictive of the words that the speaker is trying to access, and suggest that repetition is functional in helping access upcoming words. Repetition is something that the speaker would benefit from learning to do when facing a problem of retrieval, rather than a perseveration error that she should try to suppress. As mentioned earlier, this conclusion is contingent on three relatively uncontroversial assumptions: that words activate likely continuations, that repetition provides repeated words with additional activation, and that this activation spreads to the likely continuations.

Speakers appear to repeat the number of words that would be particularly helpful for accessing the future (Figure 1). Thus, the mechanisms that determine how many words are repeated result in repetitions that are particularly helpful. The functionality of repetitions helps explain why repetitions exist, and the fact that longer repetitions tend to be more helpful for accessing the future helps explain why speakers would invest the effort in reactivating the past when trying to move forward. Because repetition behavior is functional, the consequences of performing it are positive—a tip-of-the-tongue state is resolved and the speaker can continue to talk, keeping the floor and moving closer to their conversational goals. It therefore pays off to repeat, and repeating more than one word results in a higher payoff. Therefore, repetition should be reinforced by its positive consequences and grow more prevalent with experience (Skinner, 1981).

While data on this point are limited, it suggests that there is a positive correlation between repetition behavior and experience in a domain. This positive correlation distinguishes repetition from perseveration error, which is particularly rare in expert performance (Dell et al., 1997). In particular, professional actors produce many more sentence-initial repetitions than novices when trying to recall a text they have memorized (Intons-Peterson & Smyth, 1987). In second language acquisition, a higher proportion of disfluencies in proficient speakers is composed of repetitions (Derwing et al., 2004; Olynyk et al., 1990; Witton-Davies, 2010). In athletic performance, professionals are more likely to produce repetitive pre-performance routines than amateurs (Dömötör et al., 2016; MacPherson et al., 2009). Thus, in any domain that requires fast and accurate retrieval of practiced action sequences, expertise seems to come with an increasing reliance on repetition in attempting to plan or retrieve an

[3] The distribution of surprisal difference as a function of number of words repeated is also almost identical to Figure 1. We compare the two predictors below through model comparison.

inaccessible future. This positive correlation between repetition and experience is consistent with repetition being functional, a behavior that solves a real problem, and is thereby reinforced by its consequences.

LSTM is just one possible implementation of the principle that words are predicted by preceding contexts. The strong correlations in Table 2 indicate that little hinges on the particular implementation of this principle: differences in predictiveness between contexts largely reduce to differences in predictiveness between their final words. Consequently, most of the variance in LSTM association weights between contexts and words could be captured by a simple bigram model in which a word is used to predict the next word. In previous work, we have shown that simple transitional probabilities from such a bigram model can be used to predict number of words repeated (Harmon & Kapatsinski, 2015, 2016, 2020; Kapatsinski, 2005). A recurrent network has an advantage over a simple n-gram implementation that learns transitional probabilities in that it allows contexts to be arbitrarily long without running into the data sparsity issues that come from increasing the $n$ in an $n$-gram model. However, either approach implements the core principle that words become associated with predictive preceding contexts.

The simulations reported above indicate that there is a benefit to having access to the preceding context to retrieve the word that follows a disfluency ($w_1$). Even though words that follow disfluencies are relatively unpredictable to the network (Dammalapati et al., 2019), knowing the preceding word ($w_{-1}$) brings a considerable advantage, as does adding an additional preceding word ($w_{-2}$). The large reduction in surprisal from knowing $w_{-1}$ should be interpreted with caution for two reasons. First, there is widespread consensus that words are deactivated after they are produced (Dell et al., 1997). If so, then $w_{-1}$ is available to the speaker to predict $w_1$ even without repeating it, whereas preceding words ($w_{-2}$ and $w_{-3}$) are not. Thus, some of the gain from having access to $w_{-1}$ to predict $w_1$ may not require repeating $w_{-1}$, whereas all of the gain from having access to $w_{-2}$ and $w_{-3}$ requires reactivating them, which leads to the production of a multiword repetition. Second, much of the gain from knowing $w_{-1}$ is shared by $w_1$ with the words competing with it for selection, that is, its semantic competitors. Relative to semantic competitors, $w_{-2}$ appears to provide an advantage comparable to that provided by $w_{-1}$. However, it is clear that $w_{-3}$ is relatively uninformative, which aligns with the small number of three-word repetitions.

We have seen that repetitions are functional, and that the number of words repeated aligns with how many words would be most helpful to repeat (Figure 1). However, whether there is any causal relationship between how many words would be most helpful to repeat and how many words are repeated is not yet clear. The speaker is not necessarily crafting repeated chunks to maximize their effectiveness as retrieval cues for a particular word that follows. Instead, it may be that whatever mechanism produces repetitions happens to produce ones that are particularly helpful for facilitating access to the word that follows. This way of speaking is then selected over the alternatives by its consequences, growing more prevalent with experience speaking the language. Whether accessibility of the future plays an active, online role in determining how many words are repeated can only be ascertained if it can make an independent contribution to predicting repetition length. We address this question using model comparison below (Predicting Repetition Length section). Before that, a number of additional influences on repetition length need to be introduced.

## The Reactivation Hypothesis: Multiword Repetitions Require Reactivation of the Past

It is widely believed that deactivation of recently completed action units is necessary to avoid perseveration errors, i.e., unintentional repetition (Dell et al., 1997; Estes, 1972; Houghton, 1990; James, 1890; MacKay, 1982; Rumelhart & Norman, 1982). This implies that words one has already produced will need to be reactivated or re-retrieved to be re-produced, which we call the *Reactivation Hypothesis*.

We propose that the same tip-of-the-tongue state that triggers disfluency selection also triggers the speaker to attempt to reactivate the past. We hypothesize that reactivation is a process in which the past is cued, in part, by the present, that is, $w_{-1}$ cues $w_{-2}$. The present is always available to cue the past because it is the most accessible word when the speaker runs into difficulty retrieving the future: it is the word that has just been selected for execution. In contrast, $w_{-2}$ needs to be reactivated to cue $w_{-3}$.

The following context is not the only cue used to reactivate the past: as we discuss below, top-down and start cues are also involved. However, whereas these cues have the same strength of association with a particular word across utterances, the strength of the following-context cue varies. Therefore, activation from the following context should account for between-utterance differences in how easily the past can be reactivated when controlling for word identity. If reactivation of the past is needed to re-produce the past, then the strength of the association from the present to the past should predict whether a speaker produces a multiword repetition.[4]

There is evidence that listeners use this type of retrodiction to fill in words they have missed in comprehension (Gwilliams et al., 2018; Lieberman, 1963) and that they acquire backward transitional probabilities from perceptual exposure to an artificial language (Onnis & Thiessen, 2013; Pelucchi et al., 2009; Perruchet & Desaulty, 2008). The following context may also be used in planning to select modifiers and determiners that depend on the following noun context for selection. For example, the gender of a determiner depends on the following noun in German or Spanish. In English, we speak of *strong* tea but *powerful* computers, *severe* thunderstorms but *strong* hurricanes. In such cases, it appears that the modifier is selected largely based on the following context rather than the semantic differences between the alternative choices (Sinclair, 1991). We contend that backward retrieval of recently produced words is also used to reactivate the past when the future is planned but not activated enough to execute.

[4] The present is only one cue used to access the past because recently produced words are more accessible than semantically similar words that the speaker has not recently produced (Oppenheim et al., 2010), and because the words one has produced are usually repeated exactly, without semantic substitution. This high accuracy likely requires top-down cues to recently produced items that distinguish them from their semantic competitors. However, these cues are presumably either of constant strength across utterances (such as working memory slots), or are of constant strength across contexts in which an item occurs (top-down cues to the item).

## Method

To model differences in how strongly a particular following-word context activates the preceding word, we trained the LSTM backward, reversing the word order in each sentence in training. As discussed above, this training represents the learning experience that comes from guessing a word one has failed to recognize, using the following context. During test, the trained model predicted each $w_{-2}$ and $w_{-3}$ in the sample of disfluencies from the word that follows it.

## Results

Just as LSTM surprisal from the model trained forward reflects forward transitional probability, LSTM surprisal in the backward-trained model reflects backward transitional probability. This is demonstrated in Figure 2. Interestingly, the correlations for the backward-trained model are stronger than for the forward-trained model: LSTM surprisal correlates with surprisal conditional on the immediately following word at $r = 0.91$ for $w_{-2}$ and $r = 0.88$ for $w_{-3}$. These correlations are once again almost as strong as correlations between LSTM surprisal scores from LSTM models initialized with different random seeds (mean $r = .93$). That is, the probability of the backward-trained LSTM retrieving a recently produces word depends largely on the word's probability given the word that follows, backward transitional probability.

Note that backward and forward transitional probabilities do not correlate ($r = -0.16$ for $w_{-1}w_1$ bigram, $r = 0.02$ for $w_{-2}w_{-1}$ bigram, and $r = -0.14$ for $w_{-3}w_{-2}$ bigram). This means that effects of backward surprisal cannot be attributed to forward surprisal and vice versa. If number of words repeated is predictable from $p(w_{-2}|w_{-1})$, it is not predictable from $p(w_{-1}|w_{-2})$. An effect of backward surprisal is therefore diagnostic of the backward direction of processing, with $w_{-1}$ given and $w_{-2}$ being predicted.

If re-production of the past requires its retrieval using the present as a cue, we should expect that the past would be repeated only when it is probable given the present. That is, we expect a strong positive correlation between backward transitional probability and repetition length. To provide a preliminary assessment of this prediction, Figure 3 shows the density plot of backward surprisal of $w_{-2}$ and $w_{-3}$ for each repetition length. In both panels, backward surprisal is aligned with repetition length: When a word is repeated, it is more predictable (less surprising) given the following context than when it is not repeated.

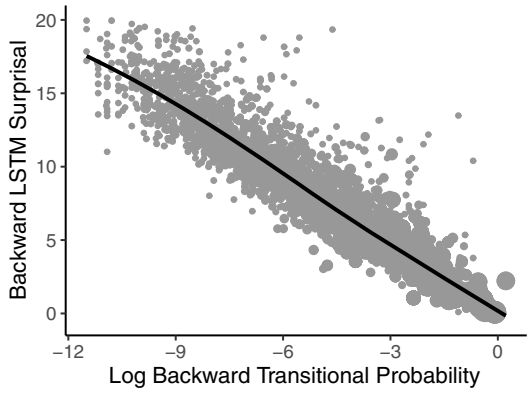
There is little correlation between backward surprisal of the past given the present and either forward surprisal of the future given the present ($r = -.05$) or the decrease in surprisal of the future that is obtained by retrieving the past ($r = -.09$). Consequently, retrievability of the future and retrievability of the past are independent influences on the length of a repetition disfluency.

## Discussion

The Reactivation Hypothesis is the contention that speakers need to re-retrieve the past to intentionally repeat it, and that they use the present as a cue in this retrieval process. It is supported by the finding that it is the backward-trained LSTM's weights that align with the number of words repeated (Figure 3). The backward weights are conditioned on the following context, corresponding to predicting the past from the present.

**Figure 2**

*The Correlation Between Backward-Trained LSTM Surprisal and Log Backward Transitional Probability*

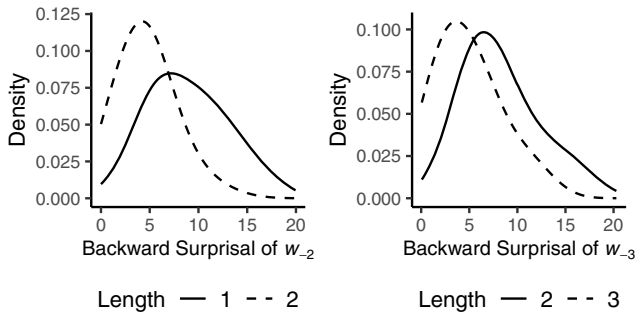



*Note*. The trendline shows a non-parametric smoother. The sizes of circles indicate the number of observations of a particular bigram. LSTM = long short-term memory.

A limitation of this analysis is that log backward surprisal is collinear with context-independent surprisal of a word, that is, negative log word frequency, with which it shares 62% of variance. Word frequency does not account for much additional variance in backward surprisal beyond what backward transitional probability accounts for (only 2% of additional variance in LSTM backward surprisal beyond backward transitional probability). Thus, LSTM surprisal largely reduces to local transitional probability. However, word frequency can provide an alternative explanation for why retrodictable words are repeated more often than words with a lower backward probability. Speakers might repeat retrodictable words because they are frequent and therefore have a higher level of resting activation across contexts (Dell, 1986; Morton, 1969), as seen from the fact that they are easier to access than rare words in picture naming (Oldfield & Wingfield, 1965). To evaluate this possibility, we will evaluate whether backward transitional probability/backward

**Figure 3**

*The Distributions of Backward-Trained LSTM Surprisal Scores for $w_{-2}$ (Left) and $w_{-3}$ (Right) when the Word is Repeated (Dashed Lines), or Not Repeated (Solid Lines)*




*Note*. Area under each curve sums to one. Repeated words tend to have lower surprisal given the words that follow them. LSTM = long short-term memory.

surprisal outperforms word frequency as a predictor of how many words are repeated below (Predicting Repetition Length section).

## The Initiation Hypothesis: Multiword Repetitions Help Reinitiate Execution

The Initiation Hypothesis proposes that certain meaningful units are better initiators than others. It is motivated by the observation that there are many morpheme boundaries in the speech stream from which production is never reinitiated. In particular, production in our corpus always restarts from a word boundary: there are no examples of reinitiating production from a word-internal morpheme or syllable boundary, as in **The cat is eating, -ing a marsupial*. This means that there are some morphemes, suffixes, that do not serve as speech initiators. This extends even to contexts where speakers make an error, because speech errors do not involve an exchange between a suffix and a prefix or stem. Restarts from suffixes are absent despite the fact that suffixes are associated with top-down semantic cues, and can serve as production units (e.g., being erroneously selected over semantically similar suffixes; Fay & Cutler, 1977; Fromkin, 1971). We propose that speakers learn not to initiate production from such units because of cue competition.

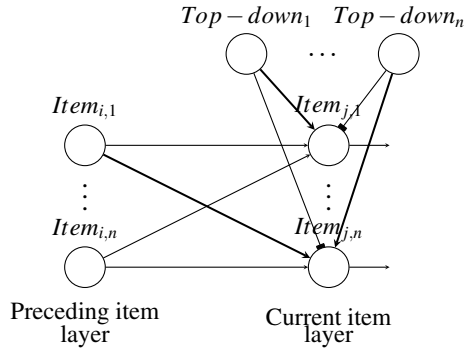
We borrow the idea of cue competition from discriminative learning theory (Rescorla & Wagner, 1972). Discriminative learning claims that co-occurring cues predicting the same outcome compete with each other for cue weight over the timecourse of learning (Arnon & Ramscar, 2012; Ramscar et al., 2010, 2013). We therefore expect cue competition between preceding-item and top-down cues: The more strongly an item is cued by the preceding item(s) in the sequence, the less strongly it will be cued by the relevant top-down cues (see also Cooper et al., 2014; Dezfouli & Balleine, 2012; Ellis, 2006; Wood & Neal, 2007).

Mechanistically, good initiators are the units that are strongly activated by whatever cues activate words in the utterance-initial position. These cues must include top-down cues—that is, semantic cues that discriminate that particular unit from other units (Figure 4; Arnon & Ramscar, 2012; Chang et al., 2006). Top-down cues are necessary to take the path less traveled, producing an item that is not the most likely item given the preceding context. That is, top-down cues exist for the purpose of overriding the influence of the preceding context when the most likely continuation is a form that does not match the speaker's intended message. In addition, we incorporated a *start cue*, present at the beginning of every utterance (Fischer-Baum & McCloskey, 2015; Henson, 1998; MacKay, 1987). In our recurrent network, this start cue simply serves as a preceding context for utterance-initial words. To the extent that words occur in that context, they can become associated with the start cue, in the same way that words become associated with preceding items they regularly follow. An utterance-boundary cue is not necessary for our model to function. However, it is useful for learning about how utterances tend to begin. For example, it allows the model to learn to initiate production with words like *well* and *so* that frequently occur in the utterance-initial position.

Competition between top-down cues and preceding context helps balance flexibility and predictability. A continuation needs to be acceptable, which means that its association with the preceding context should not be too low. That is, a word should not violate the listener's expectations. This means that activation of an upcoming item must track its predictability given the context and not be overly influenced by top-down cues. For example, the speaker should select *of* rather than *from* after *die* but *from* rather than *of* after *suffer*. Before they learn the item-to-item dependencies of English, children and non-native speakers are likely to make collocation errors by selecting the wrong continuation based on top-down, semantic input. At the same time, for the speaker to successfully produce novel utterances, top-down cues cannot be too weak. Specifically, the top-down cues to a word must be strong enough to reliably select it when the speaker decides to produce it. If a certain word is the most likely word in a particular preceding context, then its selection requires no help from top-down cues: the preceding context is sufficient to activate the word above its competitors. For example, the word *know* is highly predictable after *you*, occurring more than 30% of the time, and a large proportion of tokens of *know* are preceded by *you*. We therefore expect that *know* can be reliably produced despite having relatively weak top-down cues.

Contrast *know* with a word that is always encountered in a novel context and therefore always occurs unexpectedly. Such a word needs to have strong top-down cues because every time it is encountered the context will favor some other word, and the top-down cues will need to overcome that influence. For this reason, words that occur in a wide variety of preceding contexts should be better initiators by virtue of having stronger top-down cues. In word recognition, this idea was previously proposed by Adelman et al. (2006), who have shown that words occurring in a large number of distinct contexts are easier to recognize out of context, and introduces the term *contextual diversity* to refer to the number of distinct preceding contexts a word occurs in. Similarly, in morphological processing, morphemes that occur in a large number of distinct word contexts (i.e., morphemes of high type frequency) are easier to produce in new word contexts (Bybee, 1985). An influence of contextual diversity for maintaining top-down cue strength is also suggested by Ouellette and Wood's (1998, p.67) finding that participants who always performed some particular behavior before

**Figure 4**
*The Interaction of Top-Down and Preceding-Context Cues*




*Note*. An $item_j$ is cued by the preceding context and top-down cues. The strength of the preceding context reduces, largely, to the strength of the preceding $item_i$, which could be the preceding word or the start of the utterance. Top-down cues identify items, discriminating the *n* items in the lexicon. They are needed to overcome the influence of the preceding context when $item_j$ activated by the top-down input is not the most probable $item_j$ given $item_i$, that is, when the speaker intends to say something other than the most likely continuation of the utterance.

watching TV continued to watch TV after performing that preceding action regardless of stated goals, whereas those who performed TV viewing in a greater diversity of preceding contexts were strongly affected by their stated goals. As shown below, in LSTM, words tend to have weaker associations with preceding contexts if they occur in a wide variety of contexts. Cue competition ensures that such high-contextual-diversity words would have stronger top-down cues.

In conclusion, the stronger the preceding-context cues to a word, the less need for top-down input to select the word for production. Competition between top-down and preceding-context cues predicts that the strength of top-down cues to the word should be predictable from how unexpectedly it has occurred in the speaker's experience.

## Method

We predict that words will become better initiators whenever they occur in the initial position or in a new preceding context, and that they will be poorer initiators when they repeatedly occur medially in a familiar context, to the extent that they are predictable given the context. We implemented this hypothesis using two LSTM predictors. The first predictor is Initial Surprisal, the average surprisal of a word in the utterance-initial position in the forward-trained model, $I(w|\text{start cue})$. The second predictor is Medial Surprisal, the average surprisal of a word across non-initial positions: $\sum_{i=1}^{n} I(w_i|\text{preceding context}_i)/n$, where $n$ is the number of medial tokens of $w$. For example, if the word *book* occurs 50 times in the corpus in the non-initial position, its Medial Surprisal would be: $\sum_{i=1}^{50} I(\text{book}_i|\text{preceding context}_i)/50$.

Words with high Initial Surprisal are unexpected in the initial position. We expect these words to be poor initiators. In contrast, words with high Medial Surprisal are relatively unexpected when they occur in medial contexts. We expect these words to be good initiators. Using correlational analyses, we examined the relationship between these predictors, corpus statistics, and repetition length.

## Results

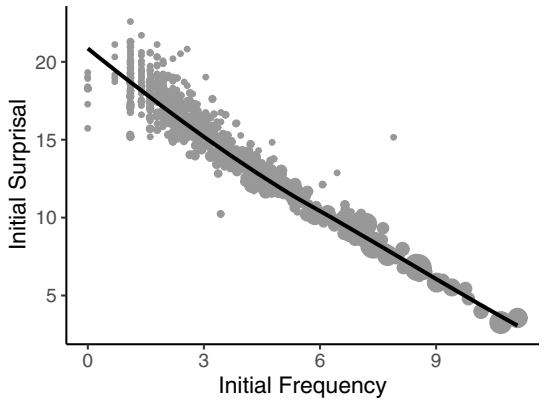
As evident from Figure 5, Initial Surprisal of a word reduces to the word's utterance-initial frequency ($r = -0.98$). That is, Initial Surprisal captures the hypothesis that words become better initiators with repeated occurrence at the beginning of the utterance.

Medial Surprisal is a more complex variable (see Table 3 and Figure 6), which increases with contextual diversity ($r = 0.75$), and decreases with token frequency in medial positions ($r = -0.86$), and mean forward transitional probability of the word ($r = -0.85$). When combined in a linear regression model, these three predictors account for 87% of the variance in Medial Surprisal, and excluding any one of them significantly reduces the fit, to 74%–76%. Thus, Medial Surprisal captures the hypothesis that medial occurrences in predictive contexts make a word a poorer initiator because a word occurring in predictive contexts can rely on these contexts for activation, making strong top-down cues to the word less necessary.

Somewhat surprisingly, Initial Surprisal and Medial Surprisal are not strongly collinear, sharing only 20% of the variance, which means they can both be entered in a regression model to predict how many words are repeated. Adding Initial Surprisal to the regression model predicting Medial Surprisal also does not improve model fit beyond the three predictors discussed above. Initial and Medial Surprisal are correlated positively: words with high Initial Surprisal also tend to have high Medial Surprisal. The shared variance is mostly explained by word frequency ($r = .91$): frequent words are less surprising in both positions.

**Figure 5**

*The Correlation Between Initial Surprisal and Log Frequency in the Utterance-Initial Position*




*Note*. The trendline shows a non-parametric smoother. The sizes of circles indicate the number of observations of a particular word.

We define a word's *Initiation Potential* as the difference between Medial Surprisal and Initial Surprisal. Initial Surprisal drives Initiation Potential down, as words that are surprising in initial positions are bad initiators, whereas Medial Surprisal drives Initiation Potential up: words that are surprising in medial positions are better initiators because their retrieval cannot rely on activation from preceding words. Figure 7 shows that a word's Initiation Potential can predict whether a speaker re-initiates production from the word. Note that the sample of disfluencies was filtered to remove all instances in which speakers actually restarted production from a clause-initial or utterance-initial word. Thus, words that tend to occur utterance-initially are better initiators even when they are clause- and utterance-medial.

Backward surprisal is negatively correlated with Initiation Potential ($r = -.36$). Though this is not a strong relationship, it means that words that are easy to reactivate also tend to be good initiators. Therefore, this correlation indicates that re-activating the past is especially likely to succeed when the past is helpful for reinitiating speech production.[5]

[5] Initiation Potential and its components are unrelated to accessibility of the future or how much it would improve from accessing the word (all $|r| \leq .1$). Thus, these are independent influences on repetition length. However, the correlation of Backward Surprisal with Initiation Potential is much weaker than its correlations with Initial Surprisal ($r = .61$) and Medial Surprisal ($r = .66$), which motivates including only the difference between the two in a regression model of repetition length. Nonetheless, we decided to include the components as predictors because we would ideally like to know whether both Initial and Medial Surprisal matter. An effect of Initial Surprisal suggests that words become better initiators when they occur in the initial position, while an effect of Medial Surprisal suggests that top-down cues weaken in proportion to the strength of preceding-context cues.

**Table 3**
*Medial Surprisal as a Function of Log Contextual Diversity, Log Medial Frequency, and Log Mean Forward Transitional Probability*

| Predictor | *b* | *SE* | *t* | *p* |
|---|---|---|---|---|
| (Intercept) | 7.14 | 0.12 | 57.34 | <.0001 |
| Contextual diversity | 0.69 | 0.02 | 32.10 | <.0001 |
| Medial frequency | −1.03 | 0.02 | −63.22 | <.0001 |
| Mean forward transitional probability | −1.02 | 0.02 | −66.79 | <.0001 |

## Discussion

The Initiation Hypothesis contends that some words are poor initiators because they co-occur with strong preceding-word cues. This prediction follows from competition between preceding-word cues and the top-down cues that are necessary to activate a word without help from a predictive preceding context, e.g., at the beginning of an utterance. These cues are top-down cues from the word's lexical/semantic representation and the start cue representing the utterance-initial context. Therefore, occurrence at the beginning of the utterance as well as in novel or unfavorable contexts should increase Initiation Potential, while occurrence in predictive contexts should decrease it. As shown in Figures 5–6, Initiation Potential corresponds to the difference between Medial Surprisal and Initial Surprisal in the LSTM model. As shown in Figure 7, a word's Initiation Potential covaries with the likelihood of reinitiating production from the word; at least for words that the speaker had already produced (panels b–d).

It is worth discussing the small or absent effect of Initiation Potential of $w_{-1}$, the word that the speaker always has access to, and uses to cue the past and the future (Figure 7a). A likely explanation is that this word is always activated enough to execute: it must be activated enough to produce because the speaker is in fact producing it at the time production stalls. Therefore, it does not need additional activation from the start cue. However, an alternative explanation is suggested by the fact that, as evident from Figure 7, the sample of words in the $w_{-1}$ position includes relatively few poor initiators. For example, there are only 64 tokens of $w_{-1}$ below Initiation Potential of −10, compared to 282 and 295 for $w_{-2}$ and $w_{-3}$ respectively. This suggests that speakers may not attempt producing a repetition at all when the word they have access to is a poor initiator: When $w_{-1}$ is a bad initiator, the utterance tends not to appear in the sample of repetitions. These explanations differ in whether the effect of Initiation Potential for $w_{-1}$ is non-significant because of small effect size or high uncertainty about effect size. We will disentangle these possibilities below using Bayesian parameter estimation.

An alternative view of Initiation Potential is to conceive of it as resulting from an error during the reactivation process. On this view, as the speaker is attempting to reactivate recently produced words by using the following words as cues, she can instead erroneously retrieve the start cue indicating the beginning of the utterance, and then restart production. This is an error in the examples we analyzed because we excluded cases in which speech was restarted from an actual utterance boundary. During reactivation, words that regularly occur utterance-initially should be better cues to the start cue. Specifically, this hypothesis proposes that the likelihood of initiating production from a word is a function of the backward surprisal of the start cue given the word. Backward surprisal of a cue is the inverse of its log backward transitional probability. Thus, backward surprisal of the start cue given a word, $I(\text{start cue}|w)$, is equivalent to the difference between a word's log frequency in initial position and its overall frequency—that is, the logarithm of the proportion of its occurrences that are utterance-initial. This measure has a strong correlation with Initiation Potential as defined above ($r = -.84$), and constitutes an alternative implementation of Initiation Potential. We will compare these two implementations below through model comparison in regression models predicting Repetition Length.

## Control Predictors: Syntax and Prosody

In this section, we introduce control predictors, which represent the current alternative explanations for why speech is restarted from certain positions and not others (Clark & Wasow, 1998; Levelt, 1983). These predictors are different in kind from the ones we considered above because they are grounded in a structuralist rather than usage-based approach to language. The usage-based framework seeks explanations for "why languages are the way they are" in the statistics of experience, picked up by domain-general learning mechanisms, as well as in the functions linguistic behaviors perform and to which they are adapted through cultural transmission (Bybee, 2001a, 2002, 2006, 2010; Christiansen & Chater, 2016; Kirby,

**Figure 6**
*Medial Surprisal as a Function of Log Contextual Diversity Controlled for Frequency (Left), Log Medial Frequency (Middle) and Log Average Forward Transitional Probability (FTP), That is, Average Probability Conditional on the Preceding Word (Right)*

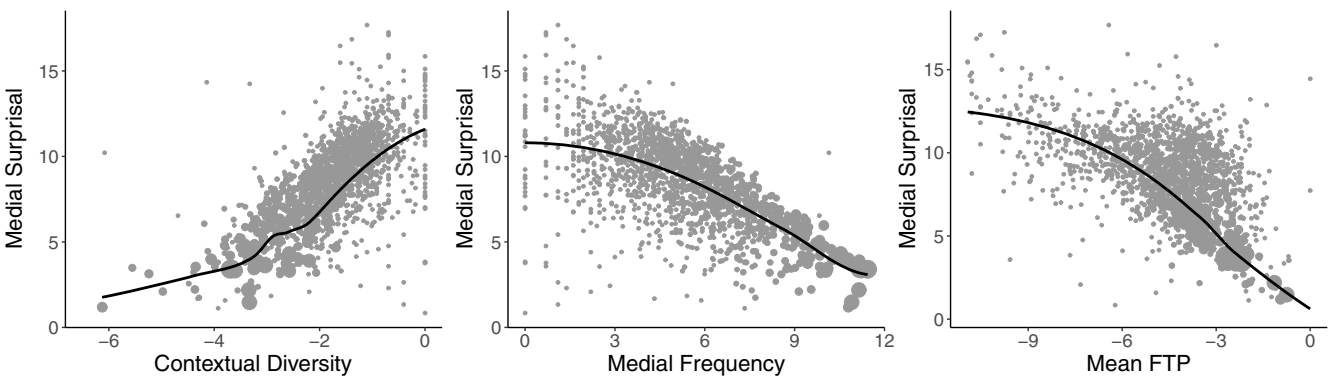
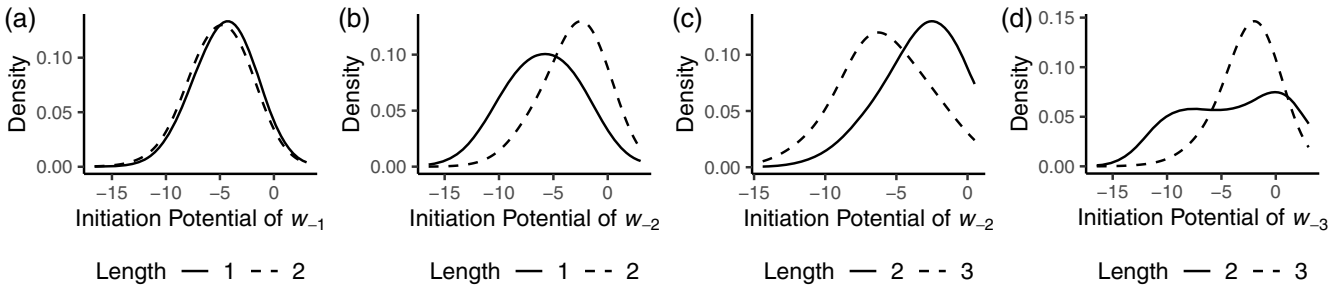


**Figure 7**

*The Distribution of Initiation Potential (Medial Surprisal − Initial Surprisal) as a Function of the Word's Position and the Number of Words Repeated (Length)*




*Note.* Panels (a) and (b) show Initiation Potential for $w_{-1}$ and $w_{-2}$ respectively for one- versus two-word repetitions. Panels (c) and (d) show Initiation Potential for $w_{-2}$ and $w_{-3}$ respectively for two- versus three-word repetitions. The dashed lines represent longer repetitions while solid lines represent shorter repetitions.



1999; Kirby et al., 2008). In keeping with this tradition, we explain repetitions with reference to the function they perform (facilitating retrieval) and the phenomenon of cue competition, which characterizes discriminative learning across domains (Ramscar et al., 2010). As discussed by Bybee and McClelland (2005), there is a natural fit between the usage-based approach to linguistic theory and connectionist models like recurrent networks, as both take a domain-general emergentist perspective on structure.

In contrast, the structuralist/generative tradition explains language behavior as a consequence of the structure of language. In particular, the predictors in this section explain repetition behavior with reference to boundaries of linguistic and prosodic constituents, without explaining where these boundaries come from. That explanation is outsourced to a universal set of boundary types defined over a universal set of lexical categories (e.g., Newmeyer, 2003). Because the boundary types are universal, explanatory of behavior and not emergent from linguistic experience with a particular language, boundaries of a certain universal type have the same strength (see Bybee, 2006, 2010 for a discussion). For example, every language is hypothesized to have verbs, and verb phrases (VPs), which include objects but not subjects, as in *The cat* [*ate a mouse*]$_{\text{VP}}$ or *We*[*'re writing a paper*]$_{\text{VP}}$. From a structuralist perspective, a verb phrase boundary is a verb phrase boundary, and behavior should treat all verb phrase boundaries alike, even though individual verbs differ in how strongly they co-occur with the preceding subject. This hypothesis is contradicted by effects of statistics of co-occurrence in experience. For example, Bybee (2002) argued that *units used together fuse together*, because frequent verbs fuse with the subjects they co-occur with (as in *are* fusing with *we* above), while less frequent subject-verb sequences like *cat ate* do not.

Maclay and Osgood (1959) proposed that speakers repeat function words because they try to restart production from a syntactic constituent boundary. However, this proposal has lost its explanatory power, as syntactic theory later posited a phrase boundary before every word (Levelt, 1983). Indeed, many syntacticians now posit a constituent boundary at every morpheme boundary (e.g., Halle & Marantz, 1993). Clark and Wasow (1998) therefore revised Maclay and Osgood's proposal to state that speech tends to be restarted from *major* constituent boundaries. They showed that single-word repetitions were most likely to occur at the beginnings of certain major constituents. Kapatsinski (2005) suggested that speech is restarted from the *nearest* major constituent boundary and showed that this hypothesis can predict how many words are repeated. However, syntactic structure does not capture all of the variance in repetition behavior. For example, speech is restarted from the beginning of a major constituent (adjunct) in Example (3) at the beginning of this article but not in (1) or (2). Furthermore, some major syntactic constituent boundaries appear to be exceptions to the hypothesis. Specifically, Fox and Jasperson (1995) showed that speech is rarely restarted from verb phrase boundaries, even though that boundary is considered to be the major break in the sentence on syntactic grounds. Our theory provides an explanation for this observation because verbs tend not to be good initiators, as they seldom occur utterance-initially. It also goes beyond this observation by predicting that some verbs (like *know*) should be especially unlikely to be used to reinitiate speech, and especially so when they follow an easily retrodictable word like *you*.

An alternative explanation for why verb phrases don't appear to behave like strong constituent boundaries in speech—even though on syntactic grounds they are expected to do so—makes reference to prosodic constituency. Prosodic constituency is thought to follow a hierarchy that is distinct from the hierarchy of syntactic constituency, hence syntactic and prosodic boundaries do not always align (Selkirk, 1984). Levelt et al. (1999) proposed that speech is restarted from the beginning of a prosodic constituent called the phonological word. This proposal accounts for why speech does not restart from verbs that are phonologically fused with the preceding subject, like *'re* in *We're writing a paper*.

The notion of a phonological word has been criticized by usage-based linguists, who consider it to be a label for sequences of words that are accessed together as a result of co-occurrence. As pointed out by Bybee (2001a, 2001b), the linguistic evidence for phonological words consists of the fact that some assimilatory phonological processes apply across some word boundaries but not others. However, the likelihood of assimilation across a word boundary is probabilistic and appears to be best explained by probabilistic measures of word co-occurrence, including forward transitional probability (Bush, 2001; Bybee, 2001b; Côté, 2013; Krug, 1998). Forward transitional probability can be interpreted as influencing how likely the second word is to be accessed before the first word is complete. That such access is required to produce cross-boundary assimilation is supported by the finding that patients who

make few anticipatory speech errors often fail to apply cross-boundary phonology (Lange et al., 2017). We therefore believe that phonological word boundaries are better considered epiphenomenal signs of co-occurrence and the resulting anticipation, rather than a distinct explanatory variable. However, because they are proposed to explain the location from which speech is initiated (Levelt et al., 1999), one could argue they should be included in an analysis as a control variable.

Unfortunately, there are no clear coding criteria for what constitutes a phonological word that can be applied independently of syntactic structure and co-occurrence (Bybee, 2001b). Thus, we use duration as a proxy for prosodic constituency because the words that adjoin to following words to form a phonological word are short function words and are thought to be further shortened by being fused into a phonological word (Selkirk, 1996). Thus, a word will be shorter when it is incorporated into a phonological word, and will be especially long if it is followed by the boundary of a prosodic constituent (Beckman & Edwards, 1990). Accordingly, if prosody accounts for the effects of experience, or at least mediates the relationship between co-occurrence and repetition behavior (Turk, 2010), the effects of experience discussed in the previous sections should disappear when competing against duration in a regression model.

## Method

### *Syntactic Constituency*

We follow the coding in Kapatsinski (2005), where syntactic constituency was argued to predict how many words are repeated. Namely, we coded the location of the *nearest* major constituent boundary, where major boundaries are those of clauses, subjects, verb/tense phrases, objects, and obliques, as illustrated in (4). We included the boundaries of adverbial phrases that are outside of the verb complex in the set of major boundaries (*has just eaten* does not have a boundary but *growling* | *loudly* does), and have assumed that coordinating conjunctions (*and, but, or*) do not belong to either of the conjuncts. The resulting constituent boundaries are consistent with the Simpler Syntax framework in syntactic theory, which was developed in part to bring syntactic theory back into agreement with traditional syntactic constituency tests (Culicover & Jackendoff, 2006). Utterances were coded blindly: information about the repeat was removed from each sentence, so that only the preceding sequence of words remained. This was done to ensure that syntactic coding was not biased by the observed location of the re-initiation point.

4. |*The cat* | *that* | *bit* | *the man* | *has just eaten* | *some kind of mouse* | *in the shadows* | *and* | *was growling* | *loudly*.

While there is no theory-independent way to code syntactic boundary strength, we believe that the present coding is close to optimal for predicting re-initiation points on the basis of syntax alone, without knowledge of word co-occurrence statistics. For example, one controversial coding choice is to consider boundaries of clauses within noun phrases to be major constituent boundaries. Thus, our syntactic coding predicts that utterances interrupted inside *that bit the man* would primarily be restarted from within that clause, rather than from the beginning of the utterance. This choice improves the accuracy of syntactic predictions, because repetitions are usually short, and because utterance-initial restarts were excluded from the data. The other controversial choice is to consider conjunctions to be outside of the conjuncts, placing a boundary between *and* and *growling*. This choice also helps syntax predict repetition lengths because the rate of conjunction repetitions is relatively low.

An alternative structural account of repetition can be derived from Levelt's (1983, 1989) theory of self-repair. Levelt examined replacement repairs, based on elicited adjective phrases (*The blue triangle, I mean, the red square* . . . ), and argued that the repair (*the red square*) and the reparandum (*the blue triangle*) need to be conjoinable with *or*. Levelt also argued that repetitions are a subtype of repair in which the replaced word is never pronounced. For the sentence in (4), Levelt's criterion would predict the boundaries in (5). These boundaries largely agree with those in (4). However, Levelt's criterion posits boundaries before lexical elements, for example, *The* | *man*, because conjunctions like *The man or woman* . . . are acceptable. Yet, these boundaries are weak based on syntactic constituency tests and are not considered major boundaries in any syntactic theory (e.g., Culicover & Jackendoff, 2006). In accordance with syntactic constituency, speech is not re-initiated from these locations in producing repetitions, just as it is not initiated from them in producing grammatical sentences (cf. *The woman is eating an apple*. vs. **Woman is eating an apple*.) Second, Levelt's criterion groups complementizers together with the preceding element, preventing repeats like *The cat that, uh, that* . . . , which are quite common in the data (cf. **The cat that bit or that scratched the man* . . . vs. *The cat that bit or scratched the man* . . . or *The cat that bit the man or the dog that attacked the woman* . . . ). We therefore consider the "major syntactic boundary" criterion above to provide syntax with the best opportunity to predict the re-initiation point.

5. *The* | *cat that* | *bit* | *the* | *man* | *was* | *eating* | *some kind of* | *mouse* | *in the* | *shadows* | *and* | *growling*.

### *Duration*

We used the hand-corrected forced-aligned version of Switchboard Corpus for extracting durations (see Deshmukh et al., 1998). This version is not coded for disfluencies, so we automatically matched the extracted hand-coded disfluencies to the forced-aligned disfluencies to obtain the duration of each word. We extracted durations of all the words preceding the interruption in our data.

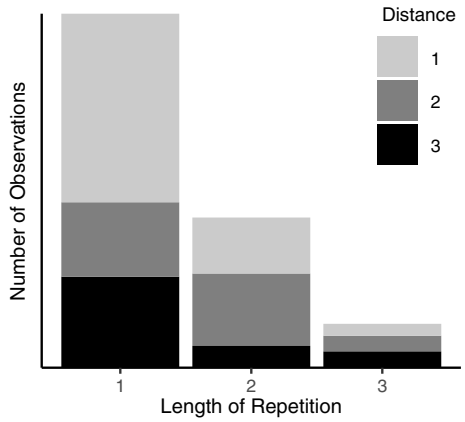
The predictors of theoretical interest for us are held responsible for accounting for variance in the location from which speech is reinitiated above and beyond these control predictors and while allowing for the existence of uncontrolled random differences between words represented by random intercepts.

## Results

Figures 8 and 9 show that the control predictors are operating as expected. There is an alignment between how many words are repeated and the location of the nearest major syntactic constituent boundary (Figure 8): Utterances in which the nearest major constituent boundary is one word away constitute a larger proportion of one-word repetitions than of two- or three-word repetitions. Similarly, utterances in which the nearest major constituent boundary is two words away constitute a larger proportion of two-word repetitions than of one-word or three-word repetitions. Repeated words

**Figure 8**

*The Relationship Between Number of Words Repeated and Distance to the Nearest Major Constituent Boundary*




*Note*. Shades show distance to the nearest major constituent boundary: Lighter means nearer.

also tend to be shorter than words that are not repeated, controlling for distance from the interruption point (Figure 9).

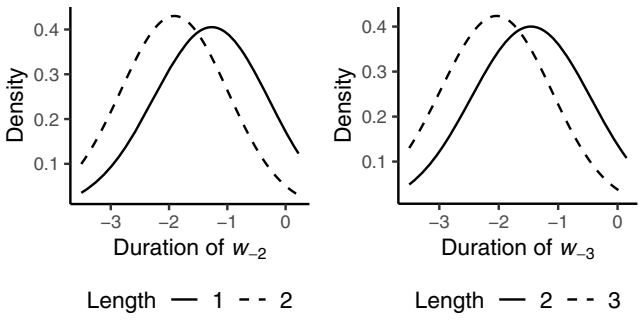
## Discussion

We have shown that the control predictors are not a strawman hypothesis: Structural factors can explain some variance in repetition length. Thus, the principles proposed by our theory will need to justify their existence by helping predict repetition length above and beyond the effects of syntax and duration.

Syntax not only predicts repetitions but also provides a reason for why repetitions tend to involve only one word: Most words preceding an interruption in speech flow are at the beginning of a major constituent. At the same time, there are many instances in which the speaker repeats more than one word even though this involves crossing the nearest major constituent boundary. Indeed, all multiword repetitions cross some phrase boundary. This shows the limitations of syntactic constituency as a predictor of repetition length.

**Figure 9**

*The Distribution of Word Duration (Prior to Repetition) as a Function of Word Position (Left: $w_{-2}$; Right: $w_{-3}$), and Whether the Word Is Repeated*




*Note*. The dashed lines represent longer repetitions while solid lines represent shorter repetitions.

Duration is intended to serve as a measurable indication of prosodic structure in the regression model below. It is, of course, an imperfect measure of prosodic structure. However, it is likely a better predictor of repetition length than a perfect measure of prosodic structure would be, because we expect the non-prosodic influences that result in shorter words to also favor repeating these words. In particular, words that are predictable from the following context are shorter than unpredictable words (Moers et al., 2017; Seyfarth, 2014), and should be more likely to be repeated due to ease of retrieval. Durations in Figure 9 are also not controlled for word identity: frequent words are shorter than rare words, as are words that happen to have more segments or syllables. More generally, words that are more difficult to produce for articulatory reasons are longer. They are also less likely to be selected for production (Martin, 2007; Schwartz & Leonard, 1982), and therefore may also be less likely to be re-produced.

## Predicting Repetition Length

While the analyses above are suggestive, they are limited because they consider one predictor at a time. Our theory, however, predicts that all of the predictors identified so far should contribute to determining how many words are repeated. Reactivation of the past should be more likely if the future is relatively inaccessible and should be easier if it is predictable from the present. Restarting from a word should also be influenced by how easy that word is to activate from its top-down semantic representation and from the utterance boundary context (start cue). This multitude of influences is typical of production choices (e.g., Bresnan et al., 2007; MacDonald, 2013; Shih, 2017), and calls for an approach that can determine whether they do all influence production choices in the predicted directions. We therefore use multiple regression to test the influence of our predictors on the length of repetitions.

## Method

All analyses reported below were conducted in R version 4.0.2 (R Core Team, 2020) using hierarchical logistic regression as implemented in the brms (Bürkner, 2017, 2018) and lme4 (Bates et al., 2015) packages. Because we used default (minimally informative) priors in brms, the maximum likelihood parameter estimates are identical between the frequentist approach implemented in lme4 and the Bayesian approach in brms. We extracted $R^2$ values using the performance package (Lüdecke et al., 2020).

### *The Dependent Variable*

Our dependent variable is the number of words repeated. We follow the approach defended by Begg and Gray (1984) in fitting two binomial models that share the reference level (two-word repetition in our case or the choice of restarting from $w_{-2}$). While brms allows for a trinomial dependent variable, the two-regression approach allows us to include different predictors and random effects for the two binary comparisons rather than including all predictors in both. This flexibility is useful because, in theory, some of the predictors (those referring to $w_{-1}$ and $w_{-3}$) are only relevant to one of the comparisons.

We conducted a sensitivity analysis to ensure that the results are robust to the choice of reference level, and choices about the inclusion of random effects and covariates. In one test, we fit a

single multinomial regression with two-word repetition as the baseline level, and all of the covariates and random effects for both contrasts. Second, we tried using a different reference level ($w_{-1}$ or $w_{-3}$) for the two regressions. Finally, we replaced modeling the choice between one-word and two-word repetition disfluencies with modeling a choice between one-word and multiword repetition disfluencies. In all cases, the results hold. Specifically, the corresponding coefficients are always on the same side of zero, and the highest posterior density (HPD) intervals that exclude zero still exclude zero.

### *Independent Variables and Model Comparison*

Model comparison is necessary to decide amongst alternative predictors that are strongly collinear if placed in the same model, and cannot be considered to jointly influence behavior. We used model comparison to decide among alternative plausible implementations of the hypotheses we proposed, and to compare them to control models.

We began with models that included only random effects and the control factors—duration of the word the speaker may or may not repeat ($w_{-2}$ for one- vs. two-word repetitions and $w_{-3}$ for two- vs. three-word repetitions), and distance to the nearest major constituent boundary. The random effects were random intercepts for $w_{-2}$, $w_{-1}$, and $w_1$ in the case of the choice to restart from $w_{-2}$ or $w_{-1}$, and the intercepts for $w_{-3}$ and $w_{-2}$ for the choice of whether to restart from $w_{-3}$ or $w_{-2}$. We excluded $w_1$ from the latter model because including it in the model resulted in a singular lme4 fit.

The models were then augmented with predictors of interest, which had to "earn their keep" by being significant, having a coefficient in the expected direction, and improving the fit of the model based on the most conservative information criterion (BIC, the Bayesian Information Criterion). The BIC provides an approximation to the Bayes Factor, which quantifies the difference in the probabilities of the models given the data assuming that the compared models have equal a priori probability (Wagenmakers, 2007). We required predictors to reduce BIC by at least seven, which is the heuristic cutoff for providing strong evidence in favor of the more complex model according to Raftery (1995). Because there are rather few three-word repetitions, we compared models on accounting for the choice of repeating one word versus two words and use the final model to explain why speakers sometimes repeat three words.

**Facilitating Access to the Future.** We first added predictors referring to accessibility of the future, which implement the hypothesis that speakers repeat more words when this would be helpful to access the upcoming word. Accessibility of the future was expected to be the weakest influence on repetition length because the speaker does not have full access to the future when deciding how many words to repeat. It was added first because it did not show strong correlations with other predictors of theoretical interest.

We compared models incorporating four different implementations of the influence of accessibility of the future on repetition length (Table 4). These implementations differ in whether the speaker is assumed to have some idea of what they have said (top vs. bottom row) and what they are about to say (left vs. right column) when estimating accessibility of the future. Predictors in the right column are surprisal values, which assume that the speaker knows what they want to say next. These implement the hypothesis that the continuation is planned but not accessible enough to be executed. Predictors on the left are entropy values, which average surprisal over all possible continuations: $H(X) = -\sum_{i=1}^{n} P(x_i|\text{preceding context}) \log P(x_i|\text{preceding context})$, where $x_i$ denotes a particular continuation and the index $i$ iterates over all possible distinct continuations, that is, word $w_1$ types. The entropy measures assume that the future has not been planned and is unknown to the speaker.

**Table 4**
*Predictors of Future Accessibility for One- vs. Two-Word Repetition*

| Knowledge State | Future unknown | Future planned |
|---|---|---|
| Past unknown | $H(x_1\|w_{-1})$ | $I(w_1\|w_{-1})$ |
| Past remembered | $H(x_1\|w_{-1}) - H(x_1\|w_{-2}w_{-1})$ | $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ |

*Note.* These predictors are a function of the knowledge state of the speaker when deciding whether to repeat $w_{-2}w_{-1}$, producing a two-word repetition, or only $w_{-1}$. $I$ = surprisal of $w_1$; $H$ = entropy (average surprisal in position 1).

The predictors in the top row implement the hypothesis that speakers retrace their steps when the current context is not a good cue to the future. It is the surprisal or entropy of the future given the context that would be repeated without restarting farther back ($w_{-1}$ for the choice of one- vs. two-word repetitions; $w_{-2}w_{-1}$ for two- vs. three-word repetitions). The predictors in the bottom row implement the hypothesis that the speaker repeats an extra word precisely when this would benefit activating the intended future. It is either the difference in surprisal, that is, relative surprisal, or the difference in entropy, that is, information gain, given a longer versus shorter context ($w_{-1}$ vs. $w_{-2}w_{-1}$ for the choice of one- vs. two-word repetitions; $w_{-3}w_{-2}w_{-1}$ vs. $w_{-2}w_{-1}$ for the choice of two- vs. three-word repetitions). We also calculated surprisal differences between $w_1$ and its nearest semantic competitor, as defined by LSA. We considered two versions of this predictor: one with an unknown past and one with a remembered past.

**Reactivation of the Past.** We retained the best implementation of accessibility of the future in the model and added predictors referring to how accessible the past is given the present. These included either (a) backward surprisal of $w_{-2}$, $I(w_{-2}|w_{-1})$, (b) forward surprisal of $w_{-2}$, $I(w_{-2}|w_{-3})$, or (c) context-independent surprisal of $w_{-2}$, $I(w_{-2})$. We expected backward surprisal to be superior to forward or context-independent surprisal, which do not implement accessibility of the past given the present.

**Initiation Potential.** Retaining the best predictor of past accessibility, we added predictors implementing Initiation Potential. These predictors implement the Initiation Hypothesis, the claim that words become good initiators if they tend to occur without a predictive preceding word, which includes utterance-initial contexts, and become poor initiators if they tend to occur in predictive preceding contexts. We created two models of Initiation Potential, which differ in how experience with a word changes the association between a word and the start cue encoding an utterance-initial context. The first model, henceforth *start-to-word* model, proposes that the connection from the start cue to a word varies as a function of how often the word follows the start cue, $p(w|\text{start cue})$. To reflect this, *start-to-word* included Initial and Medial Surprisal from forward-trained LSTM. The second model, henceforth *word-to-start*, instead assumes that the start cue will always activate a word when retrieved, but the ability of a word to retrieve the start cue varies as a function of its co-occurrence with the start cue; specifically, $p(\text{start cue}|w)$. To reflect this, *word-to-start* included average surprisal across all contexts in

which a word occurrred from forward-trained LSTM, and backward surprisal of the start cue from backward-trained LSTM.[6]

To test whether the start cue is necessary to account for the data, we created a version of the forward-trained LSTM trained on the corpus without the start cue. Forward Surprisal in this model correlated almost perfectly ($r > .99$) with Medial Surprisal in the model with a start cue, and did not capture Initial Surprisal. Therefore, an effect of Initial Surprisal on repetition length would suggest that utterance-initial words are preceded by a start cue.

Recall that Medial Surprisal correlates with a number of collinear corpus statistics, rather than reducing to a particular one (Figure 6). We used BIC model comparison to test whether the LSTM measures effectively capture the relevant statistics. If the LSTM model fully accounts for the effects of the speaker's experience with the language, it should fit the repetition data as well as or better than a model that predicts repetition behavior from the corresponding corpus statistics.

### *Testing Predictions With the Most Probable Model*

Once the most probable model was identified, we examined how the predictors comprising the model jointly influence repetition behavior by fitting the BIC-best model using a fully Bayesian approach in the brms package (Bürkner, 2018). We used two Monte Carlo Markov chains with 30,000 iterations, 10,000 of which were warmup iterations. The convergence diagnostics indicate that the chains have converged by the end of warmup. The lowest effective sample size for a parameter of interest was 10,445 independent observations. The hypotheses we evaluate are confirmed if the predictors that implement them have coefficients of the expected sign in the most probable model, and their HPD intervals exclude zero (Kruschke, 2014).

### *Model Fit*

In addition to comparing models with BIC, we report marginal pseudo-$R^2$ values of the final model and subset models that either exclude one of the hypothesized sources of influence on repetition length, or include only one such source of influence and the control predictors. These are intended to show the relative importance of the influences on repetition length (accessibility of the future, reactivation of the past, Initiation Potential, and the structural control predictors syntax and prosody). We expected that the principles we propose capture substantial variance that structural predictors do not capture. They also provide an estimate of how much variance in the behavior remains to be captured. We calculated the $R^2$ values using the *r2_bayes* function in the *performance* package (Lüdecke et al., 2020), which is the ratio of the variance in the values predicted by the fixed-effects predictors divided by the variance in predicted values plus the expected variance of the errors (Gelman et al., 2019). We report marginal $R^2$ values (rather than the higher $R^2$ values conditional on the random effects), which measure how well the model captures the variance in the data using fixed effects alone, marginalizing over the random effects of words.[7] The $R^2$ values and their corresponding HPD intervals quantify the proportion of variance explained by a particular model, but should not be used for model comparison purposes. The fact that $R^2$ does not adjust for model flexibility means that models with more predictors will always show a better fit.[8]

## Results

In this section, we first report the results of model comparisons based on BIC, beginning with the control model, which included only structural predictors, and then adding predictors of theoretical interest. Predictors referring to the Facilitation Hypothesis are added first, followed by predictors describing Reactivation of the past, followed by those implementing Initiation Potential. We then proceed to report on the effects of the predictors in the best model, and conclude by reporting how much variance in behavior is captured by each of the hypothesized influences on repetition behavior.

### *Model Comparison*

We begin by describing the control model, which does not include any predictors of theoretical interest. As summarized in Table 5, the control predictors influenced repetition behavior as expected from prior research. The presence of a major syntactic constituent boundary at $w_{-1}$ favored starting from $w_{-1}$ relative to utterances with farther-away boundaries. A boundary at $w_{-3}$ favored starting from $w_{-3}$ compared to utterances in which the nearest boundary was at $w_{-2}$. As reported in prior work (Kapatsinski, 2005), a boundary before $w_{-1}$ does not significantly influence the choice between whether to repeat $w_{-3}$. Duration of the word that the speaker considered repeating ($w_{-2}$ for one- vs. two-word repetitions; $w_{-3}$ for two- vs. three-word repetitions) influenced the likelihood of repeating it, with shorter words being more likely to be repeated. These results suggest that the control predictors capture the structural influences on repetition behavior identified in prior work.

**Facilitating Access to the Future.** We proceeded to identify the best model of one- vs. two-word repetitions by adding measures of future activation to the control model, which accounted for 7% of the variance in repetition length. The best measure of future activation was surprisal of $w_1$ given a shorter context relative to its surprisal given a longer context, Relative Surprisal or $I(w_1|w_{-1}) - I(w_1|w_{-2}w_{-1})$, which performed better than simple forward surprisal, $I(w_1|w_{-1})$; [$b(SE) = 0.44(0.06)$, $z = 7.46$, $p < .0001$; vs. $b(SE) = 0.17(0.06)$, $z = 2.75$, $p = .006$; $\Delta$BIC = 48; see Table

[6] Because mean forward Initial Surprisal in the start cue model is negative log utterance-initial frequency of the word, and backward surprisal of the start cue is the difference between log utterance-initial frequency and log utterance-medial frequency, it is not possible to combine the influences in *start-to-word* and *word-to-start* models in a superset model, as it becomes unindentifiable. Thus, we can test whether it is better to allow experience to vary the weight of the connection from the start cue to a word, or the weight of the connection from the word to the start cue, but not both.

[7] Random effects capture uncontrolled variance due to word identity, and are included to account for lack of independence between observations of the same word. Marginal $R^2$ is preferable to overall model fit, as captured by conditional $R^2$ and other measures like accuracy or the concordance score because Monte Carlo simulations show that overall model fit of a mixed-effect model is often the same whether or not a real predictor is included (Barth & Kapatsinski, 2018).

[8] The HPD intervals around the $R^2$ values also do not take into account the strong correlations between $R^2$ estimates of two models across locations in the sampling space. As the chains sampling parameter values explore the posterior, $R^2$ of the models containing those parameters varies, and the $R^2$ values of the models that share many parameters vary together. As a result, overlap in HPD intervals of $R^2$ between models underestimates the reliability of the differences in fit between models (Kruschke, 2014).

**Table 5**
*Repetition Length as a Function of Control Predictors, Syntax and Duration*

| Predictor | *b* | *SE(b)* | *z* | *p* |
|---|---|---|---|---|
| Model 1: one- versus two-word repetition | | | | |
| Intercept | −4.07 | 0.34 | −12.09 | <.0001 |
| Boundary at $w_{-2}$ | 2.21 | 0.22 | 9.92 | <.0001 |
| Boundary at $w_{-3}$ | 1.15 | 0.27 | 4.26 | <.0001 |
| Duration $w_{-2}$ | −0.82 | 0.09 | −9.25 | <.0001 |
| Model 2: two- versus three-word repetition | | | | |
| Intercept | −3.71 | 0.70 | −5.32 | <.0001 |
| Boundary at $w_{-1}$ | −0.24 | 0.42 | −0.58 | .55 |
| Boundary at $w_{-3}$ | 1.38 | 0.65 | 2.14 | .03 |
| Duration $w_{-3}$ | −1.02 | 0.18 | −5.60 | <.0001 |

*Note.* Duration was log transformed and standardized. Therefore, the intercept is for a word of average duration for the boundary location that favors a shorter disfluency.

A1]. The difference in relative surprisal between $w_1$ and its nearest LSA semantic competitor, $[I(sem.comp_1|w_{-1}) - I(sem.comp_1|w_{-2}w_{-1})] - [I(w_1|w_{-1}) - I(w_1|w_{-2}w_{-1})]$, was closest to the best measure but did not perform as well ($\Delta BIC = 35$). Only measures of relative surprisal were significant and decreased BIC compared to the control model ($\Delta BIC > 48$; Table A1 and A2). Entropy measures did not make a significant contribution to the model ($p > .8$; Table A3), and reduced model fit compared to the control model ($\Delta BIC = 8.3$; Table A3). These results suggest that the speaker repeats two words specifically when it would facilitate access to the upcoming word.

**Reactivation of the Past.** We retained relative suprisal in the model. We then examined the predictors referring to the predictability of the past. Of these predictors, predictability of $w_{-2}$ given $w_{-1}$ (Backward Surprisal) was highly significant for predicting whether the $w_{-2}$ was repeated [$b(SE) = 1.49(0.14)$, $z = 10.98$, $p < .0001$; Table A4) and improved model fit more than either context-independent or forward surprisal of $w_{-2}$ (all $\Delta BIC > 55$; Table A4). These results support the Reactivation Hypothesis, suggesting that the speaker reactivates recently produced words by cueing them, in part, by the words that follow.

**Initiation Potential.** We retained Backward Surprisal in the model and investigated what makes a word a good utterance initiator. Recall that we compared two models. Start-to-word model included mean Initial and Medial Surprisal from forward-trained LSTM. Word-to-start model included mean surprisal from forward-trained LSTM, and backward surprisal of the start cue in backward-trained LSTM.

Start-to-word model provided a better fit to the data than word-to-start model, with $\Delta BIC = 11$ (Table A5), suggesting that a good initiator is a word strongly activated by the start cue, rather than a word from which the start cue is likely to be inadvertently retrieved. However, both models had significant similarities. First, mean Forward Surprisal was significant in both models, in the expected direction—production was initiated from words with high mean Forward Surprisal—and improved model fit ($\Delta BIC > 16$; Table A5). Second, in both models, the strength of a word's association with the start cue mattered for $w_{-2}$, a word that the speaker needs to reactivate to a level sufficient for re-production, but not for $w_{-1}$. That is, the amount of activation received from the start cue matters for already produced words, which are not activated enough to reproduce. In contrast, activation received from the start cue by the word that the speaker has full access to, that is, $w_{-1}$, does not appear to affect repetition length.

**Fit Relative to Corpus Statistics.** We compared the LSTM model predictors to the corresponding corpus statistics, to determine whether the LSTM model accounts for the variance that comes from experience with words in context. Backward, Forward, and Initial Surprisal in LSTM performed about as well as the corresponding log transitional probabilities in the corpus ($\Delta BIC = 1–3$; Tables A6–A8). This is unsurprising because these LSTM measures closely track the corresponding conditional probabilities. In addition, models incorporating LSTM Medial Surprisal outperform the model with the corresponding corpus statistics (log contextual diversity, log medial frequency and mean log forward surprisal), with $\Delta BIC = 31$ for word-to-start (Table A9) and $\Delta BIC = 40$ for start-to-word (Table A10). The reason for these differences is that the LSTM models achieve a similar fit to models using corpus statistics, but with fewer predictors.

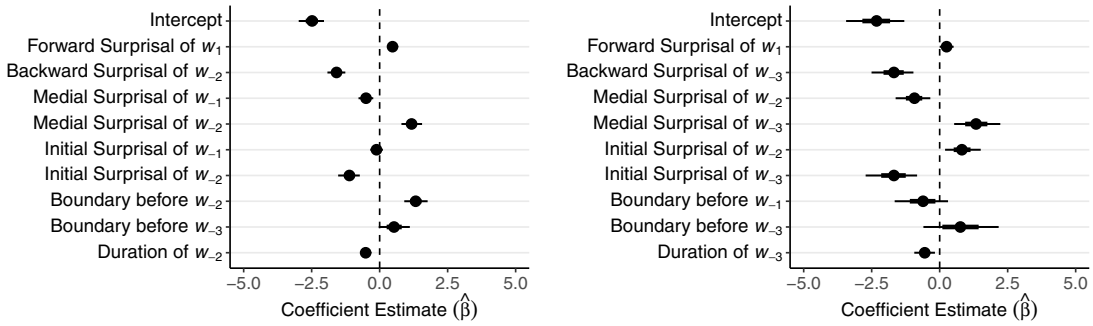
### *Coefficient Estimates in the Final Model*

The left panel of Figure 10 shows the coefficient estimates from the best (lowest-BIC) model of the choice between one-word and two-word repetitions. This was the start-to-word model. We re-implemented this model using *brms*, so that Bayesian HPD intervals for all coefficients could be derived. Intervals that span zero, marked by a vertical dashed line, indicate that zero is a believable value for the coefficient. The right panel of Figure 10 shows the results of applying the same model, with re-estimated coefficients, to the choice between repeating two versus three words. We expect the predictors of theoretical interest to have signs in the expected direction, and to have HPD intervals that do not span zero.

Overall, the results are similar for the two datasets, as shown by overlap between the corresponding HPD intervals in the right and left panels (see Table A11 for details of each model). Shorter repetitions are favored over longer ones. Repetitions are longer when repeating more would strongly boost the predictability of the future (reducing Forward Surprisal). Words are repeated when they are predictable given the word that follows (low Backward Surprisal). The speaker restarts production from words that tend to be unexpected when they occur medially (high Medial Surprisal), and expected when they occur initially (low Initial Surprisal). The significant effect of Medial Surprisal suggests the existence of competition between top-down and preceding-word cues: top-down cues to words that occur in favorable contexts are weaker than top-down cues to words that occur unpredictably. The significant effect of Initial Surprisal suggests that words that regularly occur in the initial position become associated with the utterance-initial context.

There is only one reliable difference between the two panels, as indicated by lack of overlap between 95% HPD intervals for the corresponding coefficients: Initial Surprisal does not affect the likelihood of restarting production from $w_{-1}$, the word that the speaker tends to be producing while initiating the repetition. The narrow HPD interval around the Initial Surprisal coefficient for $w_{-1}$ means that there is a reliable difference in the effect of Initial Surprisal for $w_{-1}$ versus other words. That is, the lack of effect of Initial Surprisal in

**Figure 10**
*The Final Model of Repetition Length*




*Note*. Left panel: the choice of whether to repeat one or two words (restarting from $w_{-1}$ or $w_{-2}$ respectively). Right panel: the choice of whether to repeat two or three words (restarting from $w_{-2}$ or $w_{-3}$). Positive coefficients favor longer disfluencies. 95% highest posterior density (HPD) intervals shown.

Figure 7a appears to be real, rather than being due to low variability in Initial Surprisal among $w_{-1}$'s.

Linguistic experience influences repetition behavior beyond the structural effects of syntax and duration, which are in the expected direction. Words are more likely to be repeated when they are short (Duration). The speaker also does tend to restart production from the nearest major constituent boundary: when the boundary falls before $w_{-1}$, the speaker tends not to repeat a unit that spans that boundary (left panel), and when it falls before $w_{-2}$ speakers are likely to restart speech from it. However, syntactic constituency is no longer a significant predictor of the choice between a two-word and a three-word repetition once the effects of experience are included in the model.

Figure 11 shows the proportion of variance accounted for by the full model (38%), the control model (7%), and models that exclude one or more of the hypothesized sources of influence on repetition behavior. The most important influence on the number of words repeated appears to be Initiation Potential. Of the contributors to Initiation Potential, strength of association with the start cue (Initial Surprisal) is a stronger predictor than cue competition (Medial Surprisal). The effect of Initial Surprisal means that words become better initiators when they occur utterance-initially, strengthening their association with the start cue. The effect of Medial Surprisal means that words become poorer initiators when they are frequently cued by predictive preceding contexts. The second-most-important predictor is the influence of backward associations on how easy the past is to reactivate (Backward Surprisal). Finally, the smallest influence is variability in how much repeating more words would help activating the planned future over other words (Future). Thus, the word that the speaker restarts from, and as a result the number of words repeated, is determined mostly by how strongly the candidate initiators are activated by top-down input, the start cue, and the following context.

## Discussion

In this section, we used the proposed theory to predict how many words will be repeated on each individual occasion. We have shown that predictors implementing each of the hypotheses that comprise the theory accounts for variance in repetition length beyond random or uncontrolled variation between words and the control predictors of syntax and prosody. As predicted by the Facilitation Hypothesis, words are repeated if this would help activate the future. As predicted by the Reactivation Hypothesis, words are repeated if they are strongly cued by the words that follow, as well as by the utterance-initial start cue. Finally, as predicted by the Initiation Hypothesis, speech is reinitiated from words that tend to occur in utterance-initial contexts, and tends not to be reinitiated from words that have a history of occurring in favorable preceding-word contexts (Figure 10). Together, the predictors implementing these hypotheses account for ∼35% of variance in repetition length, with control predictors improving fit by an additional 3% (Figure 11).

The proposed model appears to successfully capture variance in repetition length that is due to linguistic experience (i.e., word co-occurrence): it achieves the same fit to the data as a model that includes all of the corresponding corpus statistics, that is, frequency in utterance-initial and medial positions, forward and backward transitional probability, contextual diversity, and mean transitional probability. The BIC favors the LSTM model over the corpus model because it is able to capture the same variance in behavior using a smaller number of free parameters.[9] The model therefore appears to capture how linguistic experience influences the production of repetitions: it is sensitive to the statistics of linguistic experience in the same way speakers are.

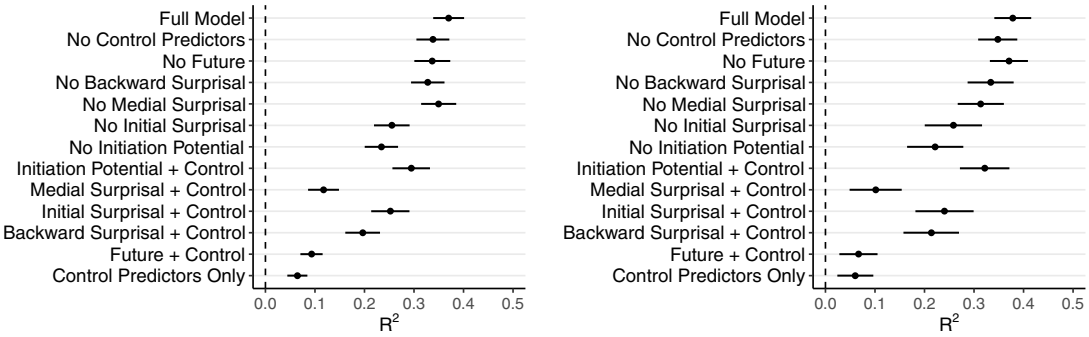
### *Facilitation*

The results for predictors of future accessibility shed some light on the information to which the speaker has access when deciding how many words to repeat. We found that the difference in surprisal given $w_{-1}$ versus $w_{-2}w_{-1}$ (relative surprisal) is the best measure of future accessibility to include in the model of repetition length. The finding that surprisal of the specific word that follows the disfluency outperforms entropy (average surprisal) suggests that the speaker has some inkling of what they need to say next when deciding to produce a multiword repetition. This result fits well with the

[9] Of course, the LSTM also has additional free parameters not seen by the BIC, such as the number of training epochs. However, these parameters were not fit to the data in the present study.

**Figure 11**

*Variance Explained by the Final Model Followed by Subset Models Excluding, in Order, Control Variables (No Control Predictors), Then One of the Hypothesized Sources of Influence: Accessibility of the Future (No Future) and the Past (No Backward Surprisal) from the Present, Followed by Medial and Initial Surprisal, the Components of Initiation Potential; Followed by Models Including Only One of These Hypothesized Sources of influence Combined With the Control Predictors of Syntax and Duration; Followed by a Model With Only Control Predictors*




hypothesis that repetition disfluencies tend to be produced when the speaker is in a tip-of-the-tongue state. That is, she has planned what to say next at the semantic level, and remains committed to this plan (Branigan et al., 1999). The intended continuation is activated more than alternatives, but is not activated enough to be selected. The repetition is then used to push the right word over the threshold. The finding that surprisal relative to the nearest LSA competitor does not perform as well indicates that the closest LSA competitor is far from the only word competing with $w_1$ for selection. The finding that relative surprisal outperforms simple forward surprisal of $w_1$ given $w_{-1}$ means that surprisal of $w_1$ given $w_{-2}w_{-1}$ matters for predicting repetition length. This suggests that an additional word is repeated specifically when it would improve access to the upcoming word.

### *Reactivation*

Backward Surprisal improves model fit more than context-independent surprisal or forward surprisal of preceding words. This result provides support for the Reactivation Hypothesis, suggesting that words are reactivated by retrodictive following contexts.

### *Initiation*

While Backward Surprisal appeared to be the strongest influence on repetition length in univariate analyses, a multiple regression approach allows us to estimate the relative contributions of the proposed influences on repetition length more precisely. The results above indicate that the strongest influence on restart location is Initiation Potential—how strongly the words preceding the interruption could be activated by cues available at the beginning of an utterance. However, the components of Initiation Potential—Initial and Medial Surprisal—do not make equal contributions. The relatively low importance of Medial Surprisal compared to Initial Surprisal contrasts with the fact that Initial Surprisal does not significantly affect $w_{-1}$, whereas Medial Surprisal significantly affects words in all three positions. We suggest that the strength of top-down input matters for words in all positions because all words receive activation from top-down input before they are reproduced. However, its effect is relatively weak because top-down cues need to be strong enough to reliably activate words across contexts, and they need to weaken enough after words are executed to prevent perseveration. These considerations place constraints on the degree to which the strength of a top-down cue to a word can be influenced by competition with preceding cues: The present always needs to be activated enough to be executed, and the past needs to be weaker than required for execution. The strength of top-down input also cannot be measured directly. All we can measure is how much it is affected by competition from context cues. As discussed below, this competition is not the only influence on the activation received from top-down cues. In particular, top-down activation must also be affected by competition for selection among the forms of semantically related words (e.g., Harmon & Kapatsinski, 2015, 2017; Harmon, 2019). As a result, the effect of cue competition on repetition length cannot be particularly strong.

The effect of Initial Surprisal is weaker for $w_{-1}$ than for other words (cf., the coefficient estimates for the sixth predictor in the two panels of Figure 10). We suggest that this weakness of the effect of Initial Surprisal for $w_{-1}$ arises because $w_{-1}$ receives more activation from top-down cues. We hypothesized that execution restarts from the word activated most highly by the cues present at the beginning of an utterance—the start cue and the top-down cues. In a neural network, the input activation of an outcome, $a$, is equal to the sum of cue activations multiplied by the weights of the associations connecting the cues to the word. Thus, $a_{\text{word}} = V_{\text{top-down}} \times a_{\text{top-down}} + V_{\text{start}} \times a_{\text{start}}$ where $V$ denotes the weight of a cue, and $a$ denotes how strongly a cue is activated.

According to this equation, the activation of a word is the sum of the two predictors we investigated, $V_{\text{top-down}}$ and $V_{\text{start}}$, weighted by the $a$ terms. We assume that the start cue is always activated equally at the beginning of an utterance, as suggested by the superior fit of start-to-word model compared to word-to-start model. This means that $a_{\text{start}}$ is constant for all words. We also assume that the top-down cues to a word are deactivated after the word is executed. That is, $a_{\text{top-down}}$ is lower for $w_{-2}$ and $w_{-3}$ than for $w_{-1}$. As a result, activation of a word that has already been executed would be affected mostly by the second term of the summation, which is proportional to the strength of the word's association with the start cue ($V_{\text{start}}$). On the other hand, activation of a word that the speaker

is currently producing will be based mostly on the first term in the summation, and therefore on the strength of its association with the top-down cues ($V_{\text{top-down}}$). The finding that the effect of $V_{\text{start}}$ is not reliably different from zero for $w_{-1}$ suggests that the contribution of top-down cues to the activation of a word selected for execution is much greater than the contribution of the start cue. The finding that the weight of the start cue is the strongest influence on the activation of $w_{-2}$ and $w_{-3}$ in turn suggests that the top-down cues to a word are deactivated after the word is executed.

### *Control Predictors*

The control predictors capture influences on repetition that are beyond the scope of our model. In particular, the effect of duration may be a side effect of differences in articulatory difficulty. There is evidence that speakers avoid attempting to produce words that are articulatorily difficult (Martin, 2007; Schwartz & Leonard, 1982). To capture this effect properly would require either feedback activation from articulatory representations to lexical selection (Martin, 2007) or a reinforcement learning model in which production difficulty resulting from selecting a word leads speakers to avoid this choice in the future. At present, we can only attempt to capture this effect by adding its acoustic correlate of duration to the regression.

The significant effect of syntactic structure on the choice between one-word and two-word repetitions may be captured in several possible ways. First, it is possible that the semantic representation of a word is also involved in reactivating the preceding word. Because words with overlapping semantics activate each other's forms (e.g., Harmon & Kapatsinski, 2015, 2017; Harmon, 2019), a word that is semantically related to the preceding word would partially activate its form. Words on different sides of a syntactic constituent boundary are relatively unrelated semantically. Thus, the semantic representations of such words have relatively little overlap, and recently produced words followed by a major boundary would be difficult to retrieve. Second, it is possible that words receive some activation from activated syntactic constituents, and that syntactic constituents become deactivated once they are executed, just as words are deactivated. In that case, words belonging to previously produced syntactic constituents would be less active than those that belong to the current one. This type of explanation requires implementing recurrent networks at multiple sequencing levels. Though this is in principle possible (Chang et al., 2006; Jordan, 1986), it is also well beyond the scope of the present work.

## General Discussion

In this article, we developed a theory of non-perseveratory repetition in language production. The proposed theory consists of three interrelated hypotheses. In this section, we revisit the evidence for each of these hypotheses, showing how they work together to motivate the existence and structure of repetitions. We further discuss how the theory explains cross-linguistic differences in reliance on repetition disfluencies, and outline directions for future work.

### The Problem of Retrieval and the Facilitation Hypothesis

The Facilitation Hypothesis motivates the occurrence of repetitions by positing that they solve the Problem of Retrieval. That is, repetitions occur when the speaker has trouble retrieving the upcoming word, and they help resolve this lexical access problem. Repeating a word increases its activation, which in turn increases the activation of the words that are predicted to occur next. We have shown that, in English, a repetition reinstates predictive cues to the words that follow and would help a recurrent neural network retrieve these words.

The model comparison results allow us to specify the Problem of Retrieval that the speaker tends to face when deploying a repetition in greater detail. We found that the speaker has some information about the upcoming word when planning a repetition: The number of words repeated depends on how well the repeated string cues the specific upcoming word (surprisal), rather than on its overall effectiveness in cueing upcoming words (entropy). This result suggests that repetitions occur when the speaker knows what to say next but has not yet settled on how to say it. Speakers plan farther ahead at the level of meaning than at the level of word form (Meyer, 1996). Repetition then appears to be deployed when execution has run ahead of lexical access at the form level, but not ahead of planning at the level of meaning (see also Blackmer & Mitton, 1991). This proposal is consistent with Branigan et al. (1999) suggestion that repetitions occur when the speaker remains committed to the current speech plan rather than trying to revise it, and contrasts with the Postma and Kolk (1993) proposal that repetitions are covert repairs.

### Multiword Repetitions and the Reactivation Hypothesis

To produce multiword repetitions, previously produced words must be reactivated. Activation-based theories of production assume that the unit selected for execution is the unit with the highest activation when that decision is made. For this reason, such theories tend to assume that production units must be deactivated after they are executed in order to avoid perseveration, that is, unintentional repetition (Dell et al., 1997; Estes, 1972; Houghton, 1990; James, 1890; MacKay, 1982; Rumelhart & Norman, 1982). Words are likely to be the relevant production units for repetition production because a repetition always restarts production from a word boundary. If words are production units, and production units are deactivated after execution, then speakers must reactivate recently produced words in order to repeat them. Without sufficient activation, the past is not repeated, even if it is accessible enough to cue the future.

We found some evidence that recently produced words remain partially accessible: the speaker is more likely to re-produce a word if this would facilitate accessing the future. As illustrated by this finding, it is functional for the speaker to retain some memory of what they have just said, in order to know whether repeating it would be useful. It is also useful to retain a memory of the past to be able to reliably re-produce it. When future is inaccessible and retrieving the past is easier than retrieving the future, retrieving the past could help access the future. Despite this, reactivation of the past is necessary to produce a multiword repetition because activation levels of recently produced units must be below the level necessary for execution in order to avoid perseveration.

We also found support for the use of backward associations in planning. We assume that the final word that the speaker is executing before the interruption point tends to be the speaker's

present, that is, the word that has not yet been deactivated. The speaker then uses this available cue, alongside other cues, to reactivate the past. We have shown that the likelihood of repeating a recently produced word is predicted by its retrievability using the following-word context, and not by its context-independent accessibility, or by its retrievability given the preceding context. This suggests that speakers learn backward associations rather than only forward ones (cf. Ramscar et al., 2010), and use these associations to reactivate the past when the future is inaccessible.

Because reactivation is fallible, multiword repetitions are uncommon relative to single-word repetitions. We hypothesized that retrieving the past involves *rewinding the chain*, so that words are retrieved one-by-one, going farther and farther into the past (Harmon & Kapatsinski, 2016; Kapatsinski, 2005; Snyder & Logan, 2014). At each point, the previously reactivated words remain available to cue the past and the future, and an additional word is added to the cueing context. Because the retrieval process moves backward word by word, words will not be skipped in producing the repetition: retrieving a word generally requires retrieving the word that follows it. The likelihood of retrieving a previously produced word is then dependent primarily on (a) its distance from the present, and (b) its backward transitional probability. Backward transitional probability matters because it is the primary determinant of how well a word is cued by the following context. Distance from the present matters because repeating the last word one has produced ($w_{-1}$) does not require retrieving anything but retrieving the second-to-last word ($w_{-2}$) requires retrieving the last word, and retrieving the third-to-last word ($w_{-3}$) requires retrieving the last word and the second-to-last word.

All of the factors described above predict repetition length beyond structural factors like syntax and prosody. Even though predictability from following context correlates with syntactic structure in a preposing language like English, backward transitional probability is not a good predictor of *all* disfluencies. In particular, Schneider (2014, 2016) shows that backward transitional probability is the worst probabilistic predictor of hesitation placement, with forward transitional probability and mutual information performing better. The proposal that repeating more than one word requires retrieval of the past using the present as a cue (the Reactivation Hypothesis), provides an explanation both for why backward transitional probability is important in accounting for the occurrence of multiword repetitions, and why it is not as important for other disfluencies, whose production does not require reactivating the past.

## The Problem of Initiation and Cue Competition

In deploying a repetition to solve the Problem of Retrieval, the speaker faces another problem, the Problem of Initiation. Repetition requires the speaker to decide from where to reinitiate execution of the speech plan. The Problem of Initiation would be trivial if speakers always produced one-word repetitions, but they do not: About 30% of repetitions in our sample involve repetition of more than one word. We proposed that the existence of multiword repetitions is motivated by the fact that a multiword context is usually more predictive of the upcoming word than a single-word context. That is, activating a longer context is more helpful for solving the Problem of Retrieval. However, having the option to produce a multiword repetition means that the speaker must find some way to decide on the location from which to reinitiate execution. We have shown that this decision is made on the basis of how strongly the words preceding the disfluency are activated by the available cues: Top-down cues representing the remembered plan, the utterance-initial context, and the words that have already been accessed.

The existence of two levels of planning (form and meaning) means that the retrieval of a word is cued by top-down input from the plan, in addition to any other cues that may be available, such as those provided by a predictive preceding context. Top-down activation is necessary to faithfully execute plans that do not correspond to the chain of most likely transitions. For example, to say *she was walking her cute little dinosaur* one would have to cue *dinosaur* by the intended meaning, as it is an unlikely continuation in the *cute little* context. In contrast, top-down control is superfluous when the word to say next is the most likely word given the preceding context: unlike *dinosaur*, *dog* would be retrievable from *the cute little* context without help from the plan. Models of production often use both top-down and preceding-context cues to choose what to do or say next (e.g., Chang et al., 2006; Cooper et al., 2014; Dell et al., 1993; Logan, 2018).

Cue competition proposes that co-occurring cues compete to predict outcomes. In the present case, top-down and preceding-context cues co-occur in cueing upcoming words, which means they can overshadow each other (Arnon & Ramscar, 2012). Based on this hypothesis, words that occur in contexts that effectively cue them should be expected to have weaker top-down cues. Cue competition is inherent to predictive models, including recurrent networks in which both top-down cues and preceding context serve to cue upcoming choices (Cooper et al., 2014; Dell et al., 1993). This is because the sum of the weights of co-occurring cues to an outcome is constant. Thus, the stronger the preceding-context cues to a word are, on average, the weaker the co-occurring top-down cues to the word.

The present results provide support for cue competition as a way of resolving the Problem of Initiation because speakers tend not to restart production from words that have occurred in predictive preceding contexts. Initiating production from a word requires activating it to a level sufficient for execution in an utterance-initial context. We suggest that this involves cueing the word with a start cue (Henson, 1998; MacKay, 1987), which occurs at the beginning of each utterance, as well as top-down cues. The top-down cues are weakened when the word occurs in a predictive preceding context. In contrast, the start cue is simply another preceding-context cue, growing stronger when the word occurs in the initial position. The strengthening of the start cue makes it easier to initiate production from words that are probable in the initial position (i.e., words that have low Initial Surprisal). Because the effect of Initial Surprisal is not captured by a model trained on a corpus in which there is no cue preceding an initial word, it provides support for the existence of a start cue. At the same time, the word's association with the start cue does not capture the effect of predictability in medial contexts, which suggests that some cue that can activate a word in the initial position weakens when the word occurs in a predictive preceding context. We suggest that this is the top-down cue to the word because top-down cues co-occur with preceding-context cues—consequently being in danger of overshadowing—and are necessary to account for how speakers navigate unlikely transitions between words.

The deactivation of recently produced words, which prevents their erroneous re-production, is naturally modeled as the withdrawal of top-down support. That is, just before a word is selected for

execution, top-down activation raises its activation to a level sufficient for the word to be executed. Once the word is executed, top-down activation flow is reduced and overall activation falls below the level necessary for execution so support from other cues is needed for the word to be re-produced. In accordance with this idea, for words that have been already produced and therefore deactivated, the activation the word receives from the start cue is the strongest influence on whether speakers will restart production from it. However, for words that have not been deactivated, the effect of activation received from the start cue is very close to zero: Speakers can start production from fully activated words even if the start cue provides no additional support.

## Why only some Languages Have Repetition Disfluencies

The Facilitation Hypothesis states that repetitions help the speaker to solve the Problem of Retrieval, cueing the future with the present. The Problem of Retrieval exists in all languages, and in other action domains: the speaker/agent constantly faces the choice of what to say/do next. It may therefore be surprising that repetitions do not exist in all languages. Specifically, Fox et al. (1996, 2009) have shown that they are not found in postposing languages like Japanese.

Both English and Japanese have easy-to-access function words and hard-to-access content words. Thus, if repetitions buy time for lexical access, one might expect that in both languages speakers could use repetition of a function word to buy time to access the upcoming content word. Why then do repetitions not occur in Japanese? The answer to this question shows how the hypotheses we proposed above fit together in explaining the emergence and maintenance of repetition behavior in a language.

Note that postposed function words are akin to suffixes. Indeed, there is no clear boundary between suffixes and postpositions because, in language change, suffixes evolve out of postpositions, fusing with the preceding items they frequently follow (Bybee, 2002; Bybee et al., 1990). One might therefore just as well ask why English speakers do not repeat suffixes when trying to access an upcoming content word. Above, we have argued that suffixes are not repeated because they are bad initiators. A suffix never occurs utterance-initially and, being a frequent unit that is required by the preceding unit in certain contexts, tends to be predictable when it occurs. For example, given *He walk*, the next morpheme in Standard English must be *-s*, or *-ed*. This is also true of postpositions: frequent, and obligatory in certain contexts, they are highly predictable from the preceding context whenever they occur (Onnis & Thiessen, 2013).

Japanese speakers could restart from a better initiator by producing a multiword repetition. However, because function words and modifiers in Japanese tend to be postposed, this would usually require reactivating a content word, a relatively difficult feat, compared to retrieving a function word or a frequent modifier like *good* or *really* that can initiate an utterance in English. That is, whereas good initiators tend to be easier to reactivate than poor initiators in English, the opposite should hold in Japanese.

Postpositions would also not facilitate access to the future as well as prepositions do: because they are semantically related to the preceding word and not the following one, they do not predict following words as well as prepositions do. Thus, repetitions in Japanese do not exist both because they would be less helpful than English repetitions for solving the Problem of Retrieval, and because they would be difficult to produce, posing significant reactivation and initiation challenges.

## Limitations and Future Directions

An important claim of the present theory is that words are cued by both context cues and top-down input from the production plan. We show that top-down input appears to be weaker for words that occur in predictable contexts. The main limitation of the present implementation is that it does not implement the top-down input cues. Variability in the strength of top-down input does not reduce to how much competition top-down cues face from contextual cues. Clearly then we are not capturing all of the variance in top-down input to a word. For example, concrete words are more accessible than abstract words (Hanley et al., 2013), likely because they are better cued by top-down input. Similarly, morphemes that express more of the intended meaning tend to be selected over more general morphemes for production (Kapatsinski, 2018a; MacWhinney, 1978). Words that have more semantic competitors are more difficult to access (Harmon & Kapatsinski, 2015; Schnadt, 2009), an effect that interacts with word frequency, as frequent words can outcompete semantically similar words for selection even when they are less semantically accurate (Harmon & Kapatsinski, 2017). An important future direction for this work is to implement top-down control of execution by distributed semantic representations.

The Initiation Hypothesis suggests that cue competition can result in top-down cues being overshadowed by context cues (Arnon & Ramscar, 2012). While generally helpful, strong context cues can lead one astray when top-down input tells one to follow a path less trodden—and can lead to an error if the top-down input is not strong enough. Wickelgren (1966) called this kind of error an *associative intrusion*. It is also known as a capture error (Norman, 1981) and a strong habit intrusion (Reason, 1979, 1990). A prediction of cue competition is that speech/language production errors arising in execution will often involve associative intrusions. A particularly good example of an associative intrusion is Benjamin Netanyahu's slip of calling the British Prime Minister *Boris Yeltsin* instead of *Boris Johnson*. because it involves the production of a rare form that is strongly cued by the preceding context. *Boris* is an exceptionally good cue to *Yeltsin* because it was almost always followed by *Yeltsin* before *Johnson*'s career took off. According to the associative account of such errors, *Yeltsin* is produced because it is activated by *Boris*, the predictive preceding item. Associative intrusions can also be exemplified in *the morals and values of my generation for the most, for most people are totally different* (the Switchboard Corpus), the speaker first produces *the* rather than *most* after *for*, a glitch that may be explained by the high predictability of *the* given a preceding *for* and the low predictability of *most* (16% vs. .03%; Corpus of Contemporary American English; Davies, 2008-2019). In *We tried it making*, *making it with gravy* (Levelt, 1989), *tried it* is over 20 times more common than *tried making*. *Tried* therefore cues *it* more strongly than it cues *making*. Levelt (1989, p. 372) notes that this is typical on *shift errors*, which tend to involve anticipation of a predictable function word. Shift errors can then also be considered to result from associative intrusion, eventually corrected with top-down input.

While sometimes denied in the literature on serial recall (Goldberg & Rapp, 2008; Henson, 1998), associative intrusion errors have received empirical support across domains. In music performance,

Chaffin and Imreh (2002) reported that expert musicians rehearsing a piece would slow down or interrupt their performance at *switchpoints*. Switchpoints are places where the musician needs to transition out of a recuerrent musical theme so that the preceding context can cue the wrong path to take. Expert musicians appear to be well aware of this difficulty of continuing down the right path and allocate additional practice time to switchpoints in rehearsing a piece. In routine action, associative intrusions have been estimated to form ~40% of all errors (Reason, 1992; Wood & Neal, 2007, p. 852; see also Norman, 1981; Reason, 1990). Speech errors have not generally been classified with associative intrusions in mind (though see Bannard et al., 2019). It is therefore unknown what proportion of errors in language production can result from priming of the error by a predictive preceding context.

In the LSTM, the average predictability of a word in context is Medial Surprisal. Medial Surprisal is necessarily correlated with contextual diversity, medial frequency, and mean surprisal conditional on the preceding word. Within a language, these are necessarily collinear, which makes it difficult to determine whether all of these influences independently affect the accessibility of a word in a new context. Controlling for word frequency, words that occur in a wide variety of contexts are less predictable in any one context. Words that occur in the same number of contexts but differ in frequency also differ in the average probability given a context. Words that have the same average probability across contexts but differ in frequency must differ in contextual diversity. However, it is possible to decouple these factors in an artificial language study by exposing speakers to production experience with systematically different mini-languages (e.g., Botvinick & Bylsma, 2005; Dell et al., 2000). Teasing apart the influences of context frequency, word frequency and contextual diversity on top-down weights is an important direction for future experimental work. For example, we could test the prediction that a word that always occurs in the same context is expected to be less accessible in a novel context than a word that occurs in a variety of contexts.

The assumption that the word before the interruption is the speaker's present, presupposes that the speaker decides to produce a repetition just before the interruption in the flow of speech. We suspect that sometimes disfluencies are preplanned because a difficulty is anticipated in advance, or the speaker wants to *perform* uncertainty for the listener (Smith & Clark, 1993). Though the empirical evidence for pre-planning in disfluency production is limited to the choice between *uh* and *um*, preplanning of repetitions may be revealed by prosodic cues, for example, how far in advance word durations begin lengthening compared to their average durations in similar environments. When a repetition is preplanned in advance, the apparent past need not be reactivated, thus we would not expect effects of Backward Surprisal or a difference in how words in different positions are affected by Medial and Initial Surprisal.

Communicative repetition, such as repetition for emphasis, is motivated by something other than a problem with lexical retrieval and instead triggered by top-down semantic input. It, therefore, falls outside the scope of the proposed theory. However, it is possible for a repetition pattern that was originally motivated by solving the problem of retrieval to become conventionalized as indicating that the speaker needs some time to think. For example, sentences like *It just . . . it just doesn't matter anymore* exhibit a conventionalized repetition, as it occurs with a surprisingly high frequency. Such repetition patterns likely have their origins in repetition disfluency—occurring when the speaker is having trouble saying what comes next—but become associated with that state of mind and can then be used communicatively to signal that what is coming up is hard to say (see also Clark & Fox Tree, 2002, for filled pauses). A repetition disfluency need not be motivated by a problem with lexical retrieval on every occasion: things can be difficult to say for other reasons. We don't necessarily expect the number of words repeated in such cases to be predictable from the predictability of the upcoming word, but the other principles we have articulated still apply. Thus, speakers are still expected to produce a multiword repetition only if the past can be reactivated and is easy to reinitiate execution from.

Speakers find it difficult to initiate word production from segments that do not occur word-initially. For example, English speakers face great difficulty pronouncing foreign words that begin with /ŋ/, like the Vietnamese name *Nguyen*, because /ŋ/ never occurs word-initially in English. Dell et al. (2000) have shown that segments can become restricted to the word-initial or postvocalic context after a short production experience in which they always occur in that position. Thus, after 20 min of producing one-syllable words that never begin with /s/, but often end in it, English speakers avoid/s/-initial productions even when making a speech error. The proposed theory suggests that production experience may also have this effect on lexical speech errors. Conversely, if competition between top-down and preceding-context cues also operates in retrieving segments constituting a word, the proposed theory also makes additional predictions for when a segment should be especially difficult to produce in an initial position. Namely, segments that tend to occur predictably should be poorer word initiators than segments that tend to occur unexpectedly.

The existence of backward associations raises the question of how speech production usually moves forward. One possibility is that backward associations are gated, and that the gates are opened only when the future is unavailable or, conversely, closed when the future is available (e.g., Sumida & Dyer, 1992). Another possibility is that backward associations are simply weaker than forward associations because there are fewer opportunities to update them. In a predictive, error-driven system, we update an association's weight only if we make a prediction error and then confirm or disconfirm it. There are fewer opportunities to retrodict the past from a present cue than to predict the future and then to confirm or disconfirm this prediction: The past is typically available before the present. While retrodiction is helpful to *fill-in* missed words, words are perceived more often than they are missed.

The proposed theory makes a number of predictions for individual differences in repetition behavior. In particular, if retrodiction is crucial for training backward associations, and backward associations in turn are essential for re-producing the past, then multiword repetition disfluencies may be especially likely—compared to single word repetitions—in speakers who must frequently use context to fill in the words they have missed, such as speakers with incipient hearing loss. This prediction highlights that the production of a multiword repetition is hypothesized to rely on additional processes that are not necessary for the production of a single-word repetition, which means that the ratio of single-word to multiword repetitions should be informative regarding the speaker's ability to remember and reactivate what they have just said. Repetitions are expected to increase with age because the growing lexicon makes the problem of retrieval more difficult (Ramscar et al., 2014), particularly so before words that are known to both older and younger speakers. More research is needed on how backward predictability changes with

experience: while upcoming words become less predictable given the upcoming context as the lexicon grows, this may not be true of the retrodiction required to produce a multiword repetition because the highly retrodictable words tend to be frequent words that are learned early. If so, then multiword repetitions may also grow more common with age and experience.

We proposed that repetition disfluencies are intentional, and are therefore categorically distinct from perseveration errors. This contrasts with the classic perspective that repetitions are something we cannot help but do whenever we have no future plans, unless we have the inhibitory control to avoid it (James, 1890). In accordance with our proposal, repetitions occur only when the speaker is *trying* to continue the utterance beyond the repeated words and therefore, do not occur at the ends of utterances. In addition, unlike perseveratory speech errors, which occur in all languages, repetition disfluencies are not universal, and occur only in languages where they would be particularly helpful for accessing the future. The rate of repetition disfluencies does not appear to correlate with inhibitory control measures (Korko & Williams, 2017) and, unlike the rate of self-corrections, is unaffected by Attention Deficit Hyperactivity Disorder (ADHD) (Engelhardt et al., 2010). Repetition disfluencies, therefore, do not appear to be something the speaker tries to suppress. However, more work is needed on the relationship between repetition and cognitive control. In particular, cognitive control may play a role in the involvement of top-down input from the production plan in the reactivation and restart process of some repetitions. For example, if more than one candidate initiation point is accessed, cognitive control processes may help decide on the location of restart.

The proposed theory rules out intentional word-final and utterance-final repetitions. However, repetitions in these positions have been reported in some individuals with autism and learning disabilities (Healey et al., 2015; Stansfield, 1995). These repetitions may be considered perseverations and thus outside the scope of the theory. However, little is known about the circumstances under which they occur, and more work is needed to determine whether they do involve a failure to deactivate a recently produced unit or are due to a more fundamental difference in how the production system is organized.

The Problem of Retrieval and the Problem of Initiation exist in all domains of action. Studying the effect of experience on how these problems are solved in the domain of language is simpler than in other domains because large corpora of the relevant behavior are available, which can provide an approximation to the speaker's experience. In other domains, this is much more difficult. Corpora of the relevant behavior are often not available. Furthermore, in some domains repetition may be covert. For example, in typing, the typist may cue upcoming words by reading the context they have typed, without retyping it, or repeating it in inner or overt speech. Repetition without re-production is possible in typing because what one has typed remain perceptually accessible to activate upcoming words without being re-produced. Covert repetition may also be useful when actual repetition cannot be performed, either because it is too motorically costly or because it would change the external physical world in an undesirable way or elicit a negative reaction from an observer. However, there is evidence that production and learning mechanisms are shared across these domains. For example, Botvinick and Plaut (2006) showed that a simple recurrent network can account for the production of action sequences, and Cooper et al. (2014) have shown that augmenting the network with top-down cues results in a better ability to stay on task. Chaffin and Imreh (2002) showed that the preceding context cues upcoming actions in piano performance and can lead the pianist astray without practiced top-down control. Allard and Starkes (1991) reported results consistent with the Initiation Hypothesis as well as the Reactivation Hypothesis in the domain of dance. They argued that modern dance differs from classical dance in that the elements of modern dance do not occur in predictive and predictable contexts, and observed that classical dancers restart production from the beginning of a routine, whereas modern dancers do not. An important direction for future work is to investigate whether the proposed hypotheses can also help explain repetition behavior in other domains of action.

Finally, one may be tempted to explain the lack of word-internal restarts by the speaker wishing to provide cues for the listener, or another observer (e.g., Hieke, 1981). Yet, we have argued that the primary function of repetition is to provide retrieval cues for the speaker. Not every motor behavior is a spectator sport, and yet the tendency to start well-practiced action sequences from the beginning, and the tendency to repeat easy actions while planning the hard actions that frequently follow them are found in every domain of skilled action. Nonetheless, it would be informative to determine whether repetitions are more likely when one is talking to an interlocutor, or is performing an action sequence for others to observe.

## Conclusion

Why do we repeat ourselves? The Facilitation Hypothesis claims that repetitions help the speaker solve the Problem of Retrieval, which arises when the upcoming word is not activated enough to be executed in time. Repetitions help resolve the resulting tip-of-the-tongue state by bombarding the likely continuation with anticipatory activation. This explains why the speaker would go to the trouble of repeating something they already said. In support of the Facilitation Hypothesis, we have shown that the repeated words are predictive cues to the words that follow.

Why then do multi-word repetitions occur far less often than single-word repetitions? It is often worthwhile to repeat more than one word because multiword repetitions are even better cues to the future than single-word repetitions. Our Reactivation Hypothesis proposes that recently produced words need to be reactivated to be re-produced, by using the present to retrodict the past. This process relies on backward associations whose direction is opposite to the flow of time.

When the problem of retrieval forces an interruption in the flow of speech, the speaker faces a Problem of Initiation. That is, the speaker must decide from where to restart execution. Speakers always restart from some word boundary, suggesting that words serve as units of execution. However, some words turn out to be better utterance initiators than others. We have argued that a word's Initiation Potential depends on how strongly it can be activated at the beginning of an utterance. This involves engaging top-down semantic cues to the word, which discriminate the word from others, as well as the utterance-initial context. We argued that top-down cues compete with co-occurring preceding-context cues. Top-down cues to a word need to be just strong enough to override the influence of a preceding context that favors continuing down a familiar but unintended path. Therefore, top-down cues are free to weaken when the word is predictable from the context, but must strengthen when it occurs unexpectedly. As expected from cue competition, speakers

are less likely to restart production from words that have occurred in favorable contexts. The Initiation Hypothesis implies that a production unit that occurs only in favorable contexts will grow increasingly restricted to such contexts and unavailable for production elsewhere, an implication that explains why speakers do not repeat suffixes or postpositions. Thus, cue competition helps account for why repetition disfluencies do not occur in some languages.

In sum, the proposed theory explains why the speaker would go to the trouble of repeating themselves while searching for the right word to say next, why some languages have repetitions and others do not, and accounts for substantial variance in how the speaker settles on the number of words to repeat on any one occasion. It improves on structural accounts of repetition, by explaining repetition behavior with reference to the speaker's linguistic experience. We hope that the proposed principles will also be tested in domains beyond language production. We suspect that production units are always cued by preceding contexts and top-down input from the plan, that these sources of information about upcoming production units compete, that agents need to reactivate the past to reproduce it, and that reactivating the past can help plan the future. Of course, only more empirical work will tell.

## References


Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Aina, L., Gulordava, K., & Boleda, G. (2019). *Putting words in context: LSTM language models and lexical ambiguity*. arXiv preprint arXiv:1906.05149.

Allard, F., & Starkes, J. L. (1991). Motor-skill experts in sports, dance, and other domains. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 126–152). Cambridge University Press.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292–305. https://doi.org/10.1016/j.cognition.2011.10.009

Bannard, C., Leriche, M., Bandmann, O., Brown, C. H., Ferracane, E., Sánchez-Ferro, Á., Obeso, J., Redgrave, P., & Stafford, T. (2019). Reduced habit-driven errors in Parkinson's Disease. *Scientific Reports*, *9*(1), Article 3423. https://doi.org/10.1038/s41598-019-39294-z

Barry, C., Hirsh, K. W., Johnston, R. A., & Williams, C. L. (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of Memory and Language*, *44*(3), 350–375. https://doi.org/10.1006/jmla.2000.2743

Barth, D., & Kapatsinski, V. (2018). Evaluating mixed-effects models of corpus-linguistic data in light of lexical diffusion. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed-effects models in linguistics* (pp. 99–116). Springer. https://doi.org/10.1007/978-3-319-69830-4_6

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beattie, G. W., & Butterworth, B. L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201–211. https://doi.org/10.1177/002383097902200301

Beckman, M. E., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. E. Beckman (Eds.), *Between the grammar and physics of speech: Papers in laboratory phonology I* (pp. 152–178). Cambridge University Press. https://doi.org/10.1017/CBO9780511627736.009

Begg, C. B., & Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, *71*(1), 11–18. https://doi.org/10.2307/2336391

Blacfkmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*(3), 173–194. https://doi.org/10.1016/0010-0277(91)90052-6

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*, 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Botvinick, M., & Bylsma, L. M. (2005). Regularization in short-term memory for serial order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 351–358. https://doi.org/10.1037/0278-7393.31.2.351

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*(2), 395–429. https://doi.org/10.1037/0033-295X.111.2.395

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, *113*(2), 201–233. https://doi.org/10.1037/0033-295X.113.2.201

Branigan, H., Lickley, R., & McKelvie, D. (1999). Non-linguistic influences on rates of disfluency in spontaneous speech. *Proceedings of the International Congress of Phonetic Sciences*, *14*, 387–389.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Royal Netherlands Academy of Arts and Sciences (KNAW).

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*(2), 204–223. https://doi.org/10.1037/0033-2909.109.2.204

Burke, D. M., Locantore, J. K., Austin, A. A., & Chae, B. (2004). Cherry pit primes Brad Pitt: Homophone priming effects on young and older adults' production of proper names. *Psychological Science*, *15*(3), 164–170. https://doi.org/10.1111/j.0956-7976.2004.01503004.x

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

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bush, N. (2001). Frequency effects and word-boundary palatalization in English. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 255–280). John Benjamins. https://doi.org/10.1075/tsl.45.14bus

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form John Benjamins*. https://doi.org/10.1075/tsl.9

Bybee, J. (2001a). *Phonology and language use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511612886

Bybee, J. (2001b). Frequency effects on French liaison. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 337–360). John Benjamins. https://doi.org/10.1075/tsl.45.17byb

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 107–134). John Benjamins. https://doi.org/10.1075/tsl.53.07byb

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *Linguistic Review*, *22*(2–4), 381–410. https://doi.org/10.1515/tlir.2005.22.2-4.381

Bybee, J. L., Pagliuca, W., & Perkins, R. D. (1990). On the asymmetries in the affixation of grammatical material. In W. A. Croft, S. Kemmer, & K. Denning (Eds.), *Studies in typology and diachrony: Papers presented*

*to Joseph H. Greenberg on his 75th birthday* (pp. 1–42). John Benjamins. https://doi.org/10.1075/tsl.20.04byb

Chaffin, R., & Imreh, G. (2002). Practicing perfection: Piano performance as expert memory. *Psychological Science*, *13*(4), 342–349. https://doi.org/10.1111/j.0956-7976.2002.00462.x

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566–581. https://doi.org/10.1037/0033-295X.103.3.566

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Clark, H. H., & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*(3), 201–242. https://doi.org/10.1006/cogp.1998.0693

Cooper, R. P., Ruh, N., & Mareschal, D. (2014). The goal circuit model: A hierarchical multi-route model of the acquisition and control of routine sequential action in humans. *Cognitive Science*, *38*(2), 244–274. https://doi.org/10.1111/cogs.12067

Côté, M. H. (2013). Understanding cohesion in French liaison. *Language Sciences*, *39*, 156–166. https://doi.org/10.1016/j.langsci.2013.02.013

Culicover, P. W., & Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends in Cognitive Sciences*, *10*(9), 413–418. https://doi.org/10.1016/j.tics.2006.07.007

Dammalapati, S., Rajkumar, R., & Agarwal, S. (2019). Expectation and locality effects in the prediction of disfluent fillers and repairs in English speech. In S. Kar, F. Nadeem, L. Burdick, G. Durrett & N.-R. Han (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 103–109. Association for Computational Linguistics.

Davies, M. (2008-2019). *The Corpus of Contemporary American English (COCA): One billion words, 1990-2019*. https://www.english-corpora.org/coca

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review*, *104*(1), 123–147. https://doi.org/10.1037/0033-295X.104.1.123

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, *17*(2), 149–195. https://doi.org/10.1207/s15516709cog1702_1

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1355–1367. https://doi.org/10.1037/0278-7393.26.6.1355

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., & Picone, J. (1998). Resegmentation of switchboard. In *Proceedings of the International Conference on Speech and Language Processing (ICSLP)* (pp. 1543–1546). Institute of Electrical and Electronics Engineers (IEEE).

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036–1051.

Dömötör, Z., Ruíz-Barquín, R., & Szabo, A. (2016). Superstitious behavior in sport: A literature review. *Scandinavian Journal of Psychology*, *57*(4), 368–382. https://doi.org/10.1111/sjop.12301

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194. https://doi.org/10.1093/applin/aml015

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Engelhardt, P. E., Corley, M., Nigg, J. T., & Ferreira, F. (2010). The role of inhibition in the production of disfluencies. *Memory & Cognition*, *38*(5), 617–628. https://doi.org/10.3758/MC.38.5.617

Estes, W. K. (1972). An associative basis for coding and organization in memory. In A. W. Mellon & E. Martin (Eds.), *Coding processes in human memory* (pp. 161–190). Winston.

Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, *8*(3), 505–520.

Ferreira, V. S., & Griffin, Z. M. (2003). Phonological influences on lexical (mis)selection. *Psychological Science*, *14*(1), 86–90. https://doi.org/10.1111/1467-9280.01424

Fischer-Baum, S., & McCloskey, M. (2015). Representation of item position in immediate serial recall: Evidence from intrusion errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1426–1446. https://doi.org/10.1037/xlm0000102

Fox, B. A., Hayashi, M., & Jasperson, R. (1996). Resources and repair: A cross-linguistic study of syntax and repair. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and Grammar* (pp. 185–237). Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.004

Fox, B. A., & Jasperson, R. (1995). A syntactic exploration of repair in English conversation. In B. P. W. Davis (Ed.), *Alternative linguistics: Descriptive and theoretical modes* (pp. 77–134). John Benjamins.

Fox, B. A., Wouk, F., Hayashi, M., Fincke, S., Tao, L., Sorjonen, M.-L., Laakso, M., & Flores Hernandez, W. (2009). A cross- linguistic investigation of the site of initiation in same-turn self-repair. In J. Sidnell (Ed.), *Conversation analysis: Comparative perspectives* (pp. 60–103). Cambridge University Press. https://doi.org/10.1017/CBO9780511635670.004

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47*(1), 27–52. https://doi.org/10.2307/412187

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). *R*-squared for Bayesian regression models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the International Conference on Acoustics, Speech, & Signal Processing (ICASSP) '92* (pp. 517–520). Institute of Electrical and Electronics Engineers.

Goldberg, A. M., & Rapp, B. (2008). Is compound chaining the serial-order mechanism of spelling? A simple recurrent network investigation. *Cognitive Neuropsychology*, *25*(2), 218–255.

Goldman-Eisler, F. (1957). Speech production and language statistics. *Nature*, *180*(4600), 1497. https://doi.org/10.1038/1801497a0

Goldman-Eisler, F. (1958). Speech production and the probability of speech in context. *The Quarterly Journal of Experimental Psychology*, *10*(2), 96–106. https://doi.org/10.1080/17470215808416261

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press.

Gulordava, K., Bojanowski, P., Grave, É., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, Vol. 1 (Long Papers)* (pp. 1195–1205). Association for Computational Linguistics.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *The Journal of Neuroscience*, *38*(35), 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In M. Halle & S. J. Keyser (Eds.), *The view from Building 20* (pp. 111–176). MIT Press.

Hanley, J. R., Hunt, R. P., Steed, D. A., & Jackman, S. (2013). Concreteness and word production. *Memory & Cognition*, *41*(3), 365–377. https://doi.org/10.3758/s13421-012-0266-5

Harmon, Z. (2019). *Accessibility, language production, and language change* [Doctoral dissertation, University of Oregon].

Harmon, Z., & Kapatsinski, V. (2015). Studying the dynamics of lexical access using disfluencies. In R. Eklund (Ed.), *Papers presented at DiSS 2015, the 7th Workshop on Disfluency in Spontaneous Speech* (pp. 41–44). University of Edinburgh.

Harmon, Z., & Kapatsinski, V. (2016). Determinants of lengths of repetition disfluencies: Probabilistic syntactic constituency in speech production. *Chicago Linguistic Society*, *50*, 237–248.

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

Harmon, Z., & Kapatsinski, V. (2020). The best-laid plans of mice and men: Competition between top-down and preceding-item cues in plan execution. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1674–1680). Cognitive Science Society.

Hartsuiker, R. J., & Notebaert, L. (2010). Lexical access problems lead to disfluencies in speech. *Experimental Psychology*, *57*(3), 169–177. https://doi.org/10.1027/1618-3169/a000021

Healey, K. T., Nelson, S., & Scott, K. S. (2015). Analysis of word-final dysfluencies in conversations of a child with autism: A treatment case study. *Procedia: Social and Behavioral Sciences*, *193*, 147–152. https://doi.org/10.1016/j.sbspro.2015.03.254

Henson, R. N. (1998). Short-term memory for serial order: The start-end model. *Cognitive Psychology*, *36*(2), 73–137. https://doi.org/10.1006/cogp.1998.0685

Hieke, A. E. (1981). A content-processing view of hesitation phenomena. *Language and Speech*, *24*(2), 147–160. https://doi.org/10.1177/002383098102400203

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In R. Dale, C. Mellish, & M. Zock (Eds.), *Current research in natural language generation* (pp. 287–319). Academic Press.

Intons-Peterson, M. J., & Smyth, M. M. (1987). The anatomy of repertory memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 490–500. https://doi.org/10.1037/0278-7393.13.3.490

James, W. (1890). *The principles of psychology*. Henry Holt.

Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach. Technical report, June 1985-March 1986* (No. AD-A-173989/5/XAB; ICS-8604). Institute for Cognitive Science, University of California.

Kapatsinski, V. (2005). Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair. *Berkeley Linguistics Society*, *30*(2), 481–92.

Kapatsinski, V. (2018a). Words versus rules (storage versus online production/processing) in morphology. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.598

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

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, *14*(4), 925–934. https://doi.org/10.1111/j.1467-7687.2011.01049.x

Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. OUP Oxford.

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

Korko, M., & Williams, S. A. (2017). Inhibitory control and the speech patterns of second language users. *British Journal of Psychology*, *108*(1), 43–72. https://doi.org/10.1111/bjop.12176

Krug, M. (1998). String frequency: A cognitive motivating factor in coalescence, language processing and linguistic change. *Journal of English Linguistics*, *26*, 286–320. https://doi.org/10.1177/007542429802600402

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lange, V. M., Cheneval, P. P., Python, G., & Laganaro, M. (2017). Contextual phonological errors and omission of obligatory liaison as a window into a reduced span of phonological encoding. *Aphasiology*, *31*(2), 201–220. https://doi.org/10.1080/02687038.2016.1176121

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38. https://doi.org/10.1017/S0140525X99001776

Levy, R. (2008). Expectation-Based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*(3), 172–187. https://doi.org/10.1177/002383096300600306

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. https://doi.org/10.1162/tacl_a_00115

Logan, G. D. (2018). Automatic control: How experts act without thinking. *Psychological Review*, *125*(4), 453–485. https://doi.org/10.1037/rev0000100

Lounsbury, F. G. (1954). Transitional probability, linguistic structure and systems of habit-family hierarchies. In C. E. Osgood & T. A. Sebeok (Eds.), *Psycholinguistics: A survey of theory and research problems* (pp. 93–101). Indiana University Press.

Lüdecke, D., Makowski, D., Waggoner, P., & Patil, I. (2020). *performance: Assessment of Regression Models Performance* (R package version 0.4.7). https://CRAN.R-project.org/package=performance

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, Article 226. https://doi.org/10.3389/fpsyg.2013.00226

MacKay, D. G. (1982). The problems of flexibility, fluency, and speed accuracy trade-off in skilled behaviors. *Psychological Review*, *89*(5), 483–506. https://doi.org/10.1037/0033-295X.89.5.483

MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills*. Springer. https://doi.org/10.1007/978-1-4612-4754-8

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*(1), 19–44. https://doi.org/10.1080/00437956.1959.11659682

MacPherson, A. C., Collins, D., & Obhi, S. S. (2009). The importance of temporal structure and rhythm for the optimum performance of motor skills: A new focus for practitioners of sport psychology. *Journal of Applied Sport Psychology*, *21*(S1), S48–S61. https://doi.org/10.1080/10413200802595930

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, *43*(1/2), 1–123.

Martin, A. (2007). *The evolving lexicon* [PhD Dissertation]. UCLA.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture–word interference experiments. *Journal of Memory and Language*, *35*(4), 477–496. https://doi.org/10.1006/jmla.1996.0026

Moers, C., Meyer, A., & Janse, E. (2017). Effects of word frequency and transitional probability on word reading durations of younger and older speakers. *Language and Speech*, *60*(2), 289–317. https://doi.org/10.1177/0023830916649215

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178. https://doi.org/10.1037/h0027366

Moss, H. E., Hare, M. L., Day, P., & Tyler, L. K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, *6*(4), 413–427. https://doi.org/10.1080/09540099408915732

Newmeyer, F. J. (2003). Grammar is grammar and usage is usage. *Language*, *79*(4), 682–707. https://doi.org/10.1353/lan.2003.0260

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1–15. https://doi.org/10.1037/0033-295X.88.1.1

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology*, *17*(4), 273–281. https://doi.org/10.1080/17470216508416445

Olynyk, M., D'Anglejan, A., & Sankoff, D. (1990). A quantitative and qualitative analysis of speech markers in the native and second language speech of bilinguals. In R. Scarcella, R. Anderson, & S. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 139–155). Newbury House.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268–284. https://doi.org/10.1016/j.cognition.2012.10.008

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227–252. https://doi.org/10.1016/j.cognition.2009.09.007

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin*, *124*(1), 54–74. https://doi.org/10.1037/0033-2909.124.1.54

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244–247. https://doi.org/10.1016/j.cognition.2009.07.011

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299–1305. https://doi.org/10.3758/MC.36.7.1299

Postma, A., & Kolk, H. (1993). The covert repair hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. *Journal of Speech, Language, and Hearing Research*, *36*(3), 472–487. https://doi.org/10.1044/jshr.3603.472

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

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

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*(1), 5–42. https://doi.org/10.1111/tops.12078

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Reason, J. T. (1979). Actions not as planned: The price of automatization. In G. Underwood & R. Stevens (Eds.), *Aspects of Consciousness* (pp. 1–61). Academic Press.

Reason, J. T. (1990). *Human error*. Cambridge University Press. https://doi.org/10.1017/CBO9781139062367

Reason, J. T. (1992). Cognitive underspecification: Its varieties and consequences. In B. J. Baars (Ed.), *Experimental slips and human error: Exploring the architecture of volition* (pp. 71–91). Plenum Press. https://doi.org/10.1007/978-1-4899-1164-3_3

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variation in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled motor performance. *Cognitive Science*, *6*(1), 1–36. https://doi.org/10.1207/s15516709cog0601_1

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*(3), 362–367. https://doi.org/10.1037/0022-3514.60.3.362

Schnadt, M. J. (2009). *Lexical influences on disfluency production* [Doctoral dissertation]. University of Edinburgh.

Schneider, U. (2014). *Frequency, chunks and hesitations. A usage-based analysis of chunking in English* [Doctoral dissertation]. University of Freiburg.

Schneider, U. (2016). Hesitation placement as evidence for chunking: A corpus-based study of spoken English. In H. Behrens & S. Pfänder (Eds.), *Experience counts: Frequency effects in language* (pp. 61–90). Mouton de Gruyter. https://doi.org/10.1515/9783110346916-004

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language*, *9*(2), 319–336. https://doi.org/10.1017/S0305000900004748

Selkirk, E. (1984). *Phonology and syntax. The relation between sound and structure*. MIT Press.

Selkirk, E. (1996). The prosodic structure of function words. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Prosodic bootstrapping from speech to grammar in early acquisition* (pp. 187–214). Lawrence Erlbaum.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155. https://doi.org/10.1016/j.cognition.2014.06.013

Shields, L. W., & Balota, D. A. (1991). Repetition and associative context effects in speech production. *Language and Speech*, *34*(1), 47–55. https://doi.org/10.1177/002383099103400103

Shih, S. S. (2017). Phonological influences in syntactic alternations. In V. Gribanova & S. S. Shih (Eds.), *The morphosyntax-phonology connection: Locality and directionality at the interface* (pp. 223–254). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190210304.003.0009

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Skinner, B. F. (1981). Selection by consequences. *Science*, *213*(4507), 501–504. https://doi.org/10.1126/science.7244649

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Snyder, K. M., & Logan, G. D. (2014). The problem of serial order in skilled typing. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1697–1717. https://doi.org/10.1037/a0037199

Stansfield, J. (1995). Word-final disfluencies in adults with learning difficulties. *Journal of Fluency Disorders*, *20*(1), 1–10. https://doi.org/10.1016/0094-730X(93)00001-V

Sumida, R. A., & Dyer, M. G. (1992). Propagation filters in PDS networks for sequencing and ambiguity resolution. *Advances in Neural Information Processing Systems*, *4*, 233–240.

Turk, A. (2010). Does prosodic constituency signal relative predictability? A smooth signal redundancy hypothesis. *Laboratory Phonology*, *1*(2), 227–262. https://doi.org/10.1515/labphon.2010.012

Turnbull, R. (2019). Listener-oriented phonetic reduction and theory of mind. *Language, Cognition and Neuroscience*, *34*(6), 747–768. https://doi.org/10.1080/23273798.2019.1579349

van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. In E. Rilloff, D. Chiang, J. Hockenmaier & Tsujii Jun'ichi (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4704–4710). Association for Computational Linguistics.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wickelgren, W. A. (1966). Associative intrusions in short-term recall. *Journal of Experimental Psychology*, *72*(6), 853–858. https://doi.org/10.1037/h0023884

Wild, F. (2015). *lsa: Latent semantic analysis* (R package version 0.73.1). https://CRAN.R-project.org/package=lsa

Witton-Davies, G. (2010). The role of repair in oral fluency and disfluency. *Selected papers from the Nineteenth International Symposium on English Teaching* (pp. 119–129). Crane.

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review*, *114*(4), 843–863.

## Appendix

This appendix provides the details of model comparisons reported in the main text. Each table reports the model summaries and Bayesian Information Criterion (BIC) values for a set of models fit to the same dataset, which allows us to compare their BIC values. Datasets differ across tables because some of the predictors are not defined for some observations in the complete dataset. For example, initial surprisal is not defined for words that have never occurred utterance-initially. As a result, BIC values are not comparable across tables. By using different datasets in different tables, we maximize the size of the dataset for each model comparison.

**Table A1**

*Repetition Length as a Function of Accessibility of $w_1$: Simple Versus Relative Surprisal*

| Predictor | *b* | *SE(b)* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Control | | | | | 3119.1 |
| (Intercept) | −3.84 | 0.30 | −12.92 | <.0001 | |
| Boundary at $w_{-2}$ | 2.13 | 0.21 | 9.93 | <.0001 | |
| Boundary at $w_{-3}$ | 1.08 | 0.26 | 4.12 | <.0001 | |
| Duration $w_{-2}$ | −0.85 | 0.09 | −9.77 | <.0001 | |
| Model 2: Forward surprisal | | | | | 3119.7 |
| (Intercept) | −3.82 | 0.30 | −12.86 | <.0001 | |
| Boundary at $w_{-2}$ | 2.11 | 0.21 | 9.85 | <.0001 | |
| Boundary at $w_{-3}$ | 1.04 | 0.26 | 3.96 | <.0001 | |
| Duration $w_{-2}$ | −0.86 | 0.09 | −9.83 | <.0001 | |
| Forward Surprisal: $I(w_1\|w_{-1})$ | 0.17 | 0.06 | 2.75 | .006 | |
| Model 3: Relative surprisal | | | | | 3067.5 |
| (Intercept) | −3.73 | 0.28 | −13.23 | <.0001 | |
| Boundary at $w_{-2}$ | 2.05 | 0.21 | 9.56 | <.0001 | |
| Boundary at $w_{-3}$ | 1.01 | 0.26 | 3.84 | .0001 | |
| Duration $w_{-2}$ | −0.84 | 0.09 | −9.64 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.44 | 0.06 | 7.46 | <.0001 | |

*Note.* BIC = Bayesian Information Criterion.

**Table A2**

*Repetition Length as a Function of Measures of Future Accessibility: $w_1$ Relative to Its Closest Semantic Competitor*

| Predictor | *b* | *SE* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Control | | | | | 3090.7 |
| (Intercept) | −3.85 | 0.30 | −12.74 | <.0001 | |
| Boundary at $w_{-2}$ | 2.11 | 0.22 | 9.77 | <.0001 | |
| Boundary at $w_{-3}$ | 1.03 | 0.26 | 3.90 | <.0001 | |
| Duration $w_{-2}$ | −0.86 | 0.09 | −9.75 | <.0001 | |
| Model 2: Relative surprisal | | | | | 3039.1 |
| (Intercept) | −3.74 | 0.29 | −13.07 | <.0001 | |
| Boundary at $w_{-2}$ | 2.04 | 0.22 | 9.41 | <.0001 | |
| Boundary at $w_{-3}$ | 0.96 | 0.27 | 3.64 | .0003 | |
| Duration $w_{-2}$ | −0.84 | 0.09 | −9.60 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.44 | 0.06 | 7.45 | <.0001 | |
| Model 3: Surprisal differences | | | | | 3093.8 |
| (Intercept) | −3.86 | 0.30 | −12.66 | <.0001 | |
| Boundary at $w_{-2}$ | 2.10 | 0.22 | 9.70 | <.0001 | |
| Boundary at $w_{-3}$ | 1.01 | 0.27 | 3.82 | .0001 | |
| Duration $w_{-2}$ | −0.86 | 0.09 | −9.73 | <.0001 | |
| $I(sem.comp_1\|w_{-1}) - I(w_1\|w_{-1})$ | −0.15 | 0.07 | −2.24 | .03 | |
| Model 4: Difference in relative surprisal | | | | | 3058.5 |
| (Intercept) | −3.80 | 0.29 | −13.11 | <.0001 | |
| Boundary at $w_{-2}$ | 2.07 | 0.22 | 9.58 | <.0001 | |
| Boundary at $w_{-3}$ | 1.01 | 0.26 | 3.81 | .0001 | |
| Duration $w_{-2}$ | −0.85 | 0.09 | −9.71 | <.0001 | |
| Difference in relative surprisal | −0.35 | 0.06 | −6.23 | <.0001 | |

*Note.* Difference in Relative Surprisal: $[I(sem.comp_1|w_{-1}) - I(sem.comp_1|w_{-2}w_{-1})] - [I(w_1|w_{-1}) - I(w_1|w_{-2}w_{-1})]$. BIC = Bayesian Information Criterion.

(*Appendix continues*)

**Table A3**

*Repetition Length as a Function of Measures of Future Accessibility: Entropy of the Future $w_1$ Relative to Its Closest Semantic Competitor*

| Predictor | *b* | *SE* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Control | | | | | 3114.1 |
| (Intercept) | −3.80 | 0.29 | −12.88 | <.0001 | |
| Boundary at $w_{-2}$ | 2.11 | 0.21 | 9.84 | <.0001 | |
| Boundary at $w_{-3}$ | 1.06 | 0.26 | 4.03 | <.0001 | |
| Duration $w_{-2}$ | −0.85 | 0.09 | −9.76 | <.0001 | |
| Model 2: Relative surprisal | | | | | 3061.6 |
| (Intercept) | −3.69 | 0.28 | −13.19 | <.0001 | |
| Boundary at $w_{-2}$ | 2.03 | 0.21 | 9.47 | <.0001 | |
| Boundary at $w_{-3}$ | 0.99 | 0.26 | 3.74 | .0002 | |
| Duration $w_{-2}$ | −0.84 | 0.09 | −9.62 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.44 | 0.06 | 7.51 | <.0001 | |
| Model 3: Entropy | | | | | 3122.4 |
| (Intercept) | −3.79 | 0.30 | −12.72 | <.0001 | |
| Boundary at $w_{-2}$ | 2.11 | 0.21 | 9.84 | <.0001 | |
| Boundary at $w_{-3}$ | 1.06 | 0.26 | 4.03 | <.0001 | |
| Duration $w_{-2}$ | −0.85 | 0.09 | −9.76 | <.0001 | |
| Entropy (Average surprisal): $H(w_1\|w_{-1})$ | 0.02 | 0.10 | 0.17 | .86 | |
| Model 4: Information gain | | | | | 3122.4 |
| (Intercept) | −3.79 | 0.30 | −12.84 | <.0001 | |
| Boundary at $w_{-2}$ | 2.11 | 0.21 | 9.84 | <.0001 | |
| Boundary at $w_{-3}$ | 1.06 | 0.26 | 4.03 | <.0001 | |
| Duration $w_{-2}$ | −0.85 | 0.09 | −9.76 | <.0001 | |
| Information gain: $H(w_1\|w_{-1}) - H(w_1\|w_{-2}w_{-1})$ | 0.02 | 0.07 | 0.23 | .82 | |

*Note.* BIC = Bayesian Information Criterion.

**Table A4**

*Model Comparison Testing Importance of Reactivation of the Past*

| Predictor | *b* | *SE* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Baseline (Relative surprisal) | | | | | 3067.5 |
| (Intercept) | −3.73 | 0.28 | −13.23 | <.0001 | |
| Boundary at $w_{-2}$ | 2.05 | 0.21 | 9.56 | <.0001 | |
| Boundary at $w_{-3}$ | 1.01 | 0.26 | 3.84 | .0001 | |
| Duration $w_{-2}$ | −0.84 | 0.09 | −9.64 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.44 | 0.06 | 7.46 | <.0001 | |
| Model 2: Backward surprisal | | | | | 2912.5 |
| (Intercept) | −3.20 | 0.25 | −12.90 | <.0001 | |
| Boundary at $w_{-2}$ | 1.51 | 0.22 | 7.00 | <.0001 | |
| Boundary at $w_{-3}$ | 0.49 | 0.27 | 1.81 | .07 | |
| Duration $w_{-2}$ | −0.52 | 0.09 | −5.83 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.47 | 0.06 | 7.94 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.49 | 0.14 | −10.98 | <.0001 | |
| Model 3: Context-independent surprisal | | | | | 3048.2 |
| (Intercept) | −3.43 | 0.26 | −13.17 | <.0001 | |
| Boundary at $w_{-2}$ | 2.04 | 0.22 | 9.45 | <.0001 | |
| Boundary at $w_{-3}$ | 1.10 | 0.27 | 4.13 | <.0001 | |
| Duration $w_{-2}$ | −0.70 | 0.09 | −7.79 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.44 | 0.06 | 7.57 | <.0001 | |
| Context-independent surprisal: $I(\mathrm{w}_{-2})$ | −0.78 | 0.16 | −4.85 | <.0001 | |
| Model 4: Forward surprisal | | | | | 3074.6 |
| (Intercept) | −3.83 | 0.31 | −12.36 | <.0001 | |
| Boundary at $w_{-2}$ | 2.06 | 0.22 | 9.51 | <.0001 | |
| Boundary at $w_{-3}$ | 1.03 | 0.27 | 3.89 | .0001 | |
| Duration $w_{-2}$ | −0.86 | 0.09 | −9.55 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.44 | 0.06 | 7.43 | <.0001 | |
| Forward surprisal: $I(w_{-2}\|w_{-3})$ | 0.11 | 0.10 | 1.11 | .269 | |

*Note.* BIC = Bayesian Information Criterion.

(*Appendix continues*)

**Table A5**
*Model Comparison Testing the Implementation of Cue Competition*

| Predictor | *b* | *SE* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Baseline (Backward surprisal) | | | | | 2701.2 |
| (Intercept) | −2.67 | 0.24 | −10.97 | <.0001 | |
| Boundary at $w_{-2}$ | 1.43 | 0.22 | 6.38 | <.0001 | |
| Boundary at $w_{-3}$ | 0.49 | 0.29 | 1.69 | .092 | |
| Duration $w_{-2}$ | −0.46 | 0.09 | −5.27 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.48 | 0.06 | 7.62 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.36 | 0.13 | −10.68 | <.0001 | |
| Model 2: Start-to-word | | | | | 2656.8 |
| (Intercept) | −2.21 | 0.22 | −9.88 | <.0001 | |
| Boundary at $w_{-2}$ | 1.29 | 0.21 | 6.05 | <.0001 | |
| Boundary at $w_{-3}$ | 0.50 | 0.28 | 1.77 | .077 | |
| Duration $w_{-2}$ | −0.47 | 0.09 | −5.42 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.47 | 0.06 | 7.74 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.29 | 0.14 | −9.47 | <.0001 | |
| Medial surprisal $w_{-1}$ | −0.44 | 0.12 | −3.76 | .0002 | |
| Medial surprisal $w_{-2}$ | 0.95 | 0.15 | 6.19 | <.0001 | |
| Initial surprisal $w_{-1}$: $I(w_{-1}\|\text{start cue})$ | −0.11 | 0.11 | −0.99 | .321 | |
| Initial surprisal $w_{-2}$: $I(w_{-2}\|\text{start cue})$ | −1.08 | 0.19 | −5.59 | <.0001 | |
| Model 3: Word-to-start | | | | | 2667.4 |
| (Intercept) | −2.32 | 0.23 | −10.11 | <.0001 | |
| Boundary at $w_{-2}$ | 1.25 | 0.21 | 5.89 | <.0001 | |
| Boundary at $w_{-3}$ | 0.41 | 0.28 | 1.46 | .145 | |
| Duration $w_{-2}$ | −0.50 | 0.09 | −5.75 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.48 | 0.06 | 7.77 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.35 | 0.13 | −10.02 | <.0001 | |
| Mean forward surprisal $w_{-1}$ | −0.52 | 0.10 | −5.04 | <.0001 | |
| Mean forward surprisal $w_{-2}$ | 0.36 | 0.15 | 2.42 | .016 | |
| Backward surprisal of start cue $I(\text{start cue}\|w_{-1})$ | −0.05 | 0.11 | −0.43 | .668 | |
| Backward surprisal of start cue $I(\text{start cue}\|w_{-2})$ | −0.70 | 0.16 | −4.47 | <.0001 | |
| Model 4: Initial surprisal | | | | | 2692.9 |
| (Intercept) | −2.21 | 0.24 | −9.23 | <.0001 | |
| Boundary at $w_{-2}$ | 1.53 | 0.23 | 6.76 | <.0001 | |
| Boundary at $w_{-3}$ | 0.68 | 0.29 | 2.30 | .021 | |
| Duration $w_{-2}$ | −0.40 | 0.09 | −4.53 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.48 | 0.06 | 7.65 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.27 | 0.14 | −9.35 | <.0001 | |
| Initial surprisal $w_{-1}$: $I(w_{-1}\|\text{start cue})$ | −0.36 | 0.10 | −3.74 | .0002 | |
| Initial surprisal $w_{-2}$: $I(w_{-2}\|\text{start cue})$ | −0.63 | 0.20 | −3.20 | .0014 | |
| Model 5: Backward surprisal of start cue | | | | | 2683.2 |
| (Intercept) | −2.62 | 0.23 | −11.48 | <.0001 | |
| Boundary at $w_{-2}$ | 1.35 | 0.22 | 6.25 | <.0001 | |
| Boundary at $w_{-3}$ | 0.48 | 0.29 | 1.68 | .093 | |
| Duration $w_{-2}$ | −0.44 | 0.09 | −5.20 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.47 | 0.06 | 7.63 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.26 | 0.13 | −9.94 | <.0001 | |
| Backward surprisal of start cue $I(\text{start cue}\|w_{-1})$ | 0.14 | 0.11 | 1.25 | .21 | |
| Backward surprisal of start cue $I(\text{start cue}\|w_{-2})$ | −0.83 | 0.15 | −5.56 | <.0001 | |

*Note.* Mean Forward Surprisal of a particular word type is its Forward Surprisal averaged over all the tokens of the word: $\sum_{i=1}^{n} I(w_{-1}|w_{-2})/n$, and $\sum_{i=1}^{n} I(w_{-2}|w_{-3})/n$, where n is the number of tokens of $w_{-1}$ and $w_{-2}$ respectively. BIC = Bayesian Information Criterion.

(*Appendix continues*)

**Table A6**
*Statistical Versus LSTM Measures of Future Predictability*

| Predictor | *b* | *SE* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Corpus future | | | | | 3053.5 |
| (Intercept) | −4.05 | 0.34 | −12.01 | <.0001 | |
| Boundary at $w_{-2}$ | 2.21 | 0.22 | 9.90 | <.0001 | |
| Boundary at $w_{-3}$ | 1.14 | 0.27 | 4.20 | <.0001 | |
| Duration $w_{-2}$ | −0.84 | 0.09 | −9.32 | <.0001 | |
| Log Forward transitional probability: log $p(w_1\|w_{-1})$ | −0.11 | 0.07 | −1.61 | .107 | |
| Model 2: LSTM future | | | | | 3051.7 |
| (Intercept) | −4.05 | 0.34 | −12.08 | <.0001 | |
| Boundary at $w_{-2}$ | 2.20 | 0.22 | 9.86 | <.0001 | |
| Boundary at $w_{-3}$ | 1.12 | 0.27 | 4.13 | <.0001 | |
| Duration $w_{-2}$ | −0.83 | 0.09 | −9.31 | <.0001 | |
| Forward surprisal: $I(w_1\|w_{-1})$ | 0.13 | 0.06 | 2.10 | .036 | |

*Note.* BIC = Bayesian Information Criterion; LSTM = long short-term memory.

**Table A7**
*Statistical Versus LSTM Measures of Past Reactivation*

| Predictor | *b* | *SE* | *z* | *p* | *BIC* |
|---|---|---|---|---|---|
| Model 1: Corpus past | | | | | 2895.4 |
| (Intercept) | −3.21 | 0.24 | −13.39 | <.0001 | |
| Boundary at $w_{-2}$ | 1.38 | 0.21 | 6.49 | <.0001 | |
| Boundary at $w_{-3}$ | 0.40 | 0.27 | 1.50 | .133 | |
| Duration $w_{-2}$ | −0.52 | 0.09 | −5.92 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.44 | 0.06 | 7.58 | <.0001 | |
| Log backward transitional probability: log $p(w_{-2}\|w_{-1})$ | 1.45 | 0.13 | 11.08 | <.0001 | |
| Model 2: LSTM past | | | | | 2894.3 |
| (Intercept) | −3.19 | 0.24 | −13.07 | <.0001 | |
| Boundary at $w_{-2}$ | 1.50 | 0.21 | 6.99 | <.0001 | |
| Boundary at $w_{-3}$ | 0.46 | 0.27 | 1.71 | .087 | |
| Duration $w_{-2}$ | −0.52 | 0.09 | −5.79 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.47 | 0.06 | 7.95 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.45 | 0.13 | −10.88 | <.0001 | |

*Note.* BIC = Bayesian Information Criterion; LSTM = long short-term memory.

**Table A8**
*Statistical Versus LSTM Measures of Initialness*

| Predictor | *b* | *SE* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Corpus initialness | | | | | 2690.1 |
| (Intercept) | −2.17 | 0.24 | −9.11 | <.0001 | |
| Boundary at $w_{-2}$ | 1.54 | 0.23 | 6.78 | <.0001 | |
| Boundary at $w_{-3}$ | 0.69 | 0.29 | 2.33 | .02 | |
| Duration $w_{-2}$ | −0.39 | 0.09 | −4.42 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.48 | 0.06 | 7.66 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.25 | 0.14 | −9.22 | <.0001 | |
| Log initial frequency $w_{-1}$ | 0.36 | 0.10 | 3.56 | .0004 | |
| Log initial frequency $w_{-2}$ | 0.75 | 0.20 | 3.76 | .0002 | |
| Model 2 LSTM initialness | | | | | 2692.9 |
| (Intercept) | −2.21 | 0.24 | −9.23 | <.0001 | |
| Boundary at $w_{-2}$ | 1.53 | 0.23 | 6.76 | <.0001 | |
| Boundary at $w_{-3}$ | 0.68 | 0.29 | 2.30 | .021 | |
| Duration $w_{-2}$ | −0.40 | 0.09 | −4.53 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.48 | 0.06 | 7.65 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.27 | 0.14 | −9.35 | <.0001 | |
| Initial surprisal $w_{-1}$: $I(w_{-1}$\|start cue) | −0.36 | 0.10 | −3.74 | .0002 | |
| Initial Surprisal $w_{-2}$: $I(w_{-2}$\|start cue) | −0.63 | 0.20 | −3.20 | .0014 | |

*Note.* BIC = Bayesian Information Criterion; LSTM = long short-term memory.

(*Appendix continues*)

**Table A9**

*Statistical Versus Word-to-start LSTM Measures of Medial Surprisal*

| Predictor | *b* | *SE* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Corpus medial surprisal | | | | | 2903 |
| (Intercept) | −2.90 | 0.23 | −12.47 | <.0001 | |
| Boundary at $w_{-2}$ | 1.43 | 0.21 | 6.80 | <.0001 | |
| Boundary at $w_{-3}$ | 0.57 | 0.27 | 2.11 | .035 | |
| Duration $w_{-2}$ | −0.52 | 0.09 | −5.68 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.46 | 0.06 | 7.96 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.37 | 0.15 | −9.14 | <.0001 | |
| Medial frequency $w_{-1}$ | 0.52 | 0.24 | 2.16 | .031 | |
| Medial frequency $w_{-2}$ | −0.38 | 0.44 | −0.88 | .38 | |
| Contextual diversity $w_{-1}$ | 0.01 | 0.27 | 0.03 | .973 | |
| Contextual diversity $w_{-2}$ | 0.64 | 0.47 | 1.37 | .17 | |
| Mean FTP $w_{-1}$ | −0.14 | 0.14 | −1.00 | .316 | |
| Mean FTP $w_{-2}$ | −0.53 | 0.20 | −2.62 | <.01 | |
| Backward surprisal of start cue: $I(\text{start cue}\|w_{-2})$ | −0.81 | 0.17 | −4.75 | <.0001 | |
| Backward surprisal of start cue: $I(\text{start cue}\|w_{-1})$ | 0.24 | 0.11 | 2.17 | .03 | |
| Model 2: LSTM medial surprisal | | | | | 2871.7 |
| (Intercept) | −2.90 | 0.23 | −12.57 | <.0001 | |
| Boundary at $w_{-2}$ | 1.33 | 0.21 | 6.47 | <.0001 | |
| Boundary at $w_{-3}$ | 0.44 | 0.26 | 1.65 | .098 | |
| Duration $w_{-2}$ | −0.56 | 0.09 | −6.18 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.47 | 0.06 | 8.12 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.44 | 0.15 | −9.91 | <.0001 | |
| Backward surprisal of the start cue $I(\text{start cue}\|w_{-2})$ | −0.83 | 0.16 | −5.22 | <.0001 | |
| Backward surprisal of the start cue $I(\text{start cue}\|w_{-1})$ | 0.03 | 0.10 | 0.35 | .729 | |
| Mean forward surprisal $w_{-1}$ | −0.43 | 0.09 | −4.51 | <.0001 | |
| Mean forward surprisal $w_{-2}$ | 0.43 | 0.15 | 2.80 | .005 | |

*Note.* BIC = Bayesian Information Criterion; LSTM = long short-term memory; FTP = Forward Transitional Probability.

**Table A10**

*Statistical Versus Start-to-word LSTM Measures of Medial Surprisal*

| Predictor | *b* | *SE* | *z* | *p* | BIC |
|---|---|---|---|---|---|
| Model 1: Corpus final | | | | | 2696.7 |
| (Intercept) | −2.38 | 0.24 | −9.97 | <.0001 | |
| Boundary at $w_{-2}$ | 1.43 | 0.22 | 6.52 | <.0001 | |
| Boundary at $w_{-3}$ | 0.62 | 0.29 | 2.18 | .029 | |
| Duration $w_{-2}$ | −0.48 | 0.09 | −5.15 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.46 | 0.06 | 7.67 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.55 | 0.16 | −9.45 | <.0001 | |
| Medial frequency $w_{-2}$ | 0.88 | 0.32 | 2.77 | .006 | |
| Medial frequency $w_{-2}$ | −1.32 | 0.50 | −2.63 | .009 | |
| Contextual diversity $w_{-1}$ | −0.10 | 0.26 | −0.39 | .695 | |
| Contextual diversity $w_{-2}$ | 0.84 | 0.48 | 1.74 | .081 | |
| Mean FTP $w_{-1}$ | −0.06 | 0.18 | −0.35 | .727 | |
| Mean FTP $w_{-2}$ | −0.89 | 0.28 | −3.23 | .0012 | |
| Initial surprisal $w_{-1}$: $I(w_{-1}\|\text{start cue})$ | 0.10 | 0.17 | 0.59 | .556 | |
| Initial surprisal $w_{-2}$: $I(w_{-2}\|\text{start cue})$ | −1.07 | 0.28 | −3.79 | .0002 | |
| Model 2: LSTM final | | | | | 2656.8 |
| (Intercept) | −2.42 | 0.22 | −10.90 | <.0001 | |
| Boundary at $w_{-2}$ | 1.29 | 0.21 | 6.05 | <.0001 | |
| Boundary at $w_{-3}$ | 0.50 | 0.28 | 1.77 | .077 | |
| Duration $w_{-2}$ | −0.51 | 0.09 | −5.42 | <.0001 | |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.46 | 0.06 | 7.74 | <.0001 | |
| Backward surprisal: $I(w_{-2}\|w_{-1})$ | −1.53 | 0.16 | −9.47 | <.0001 | |
| Medial surprisal $w_{-1}$ | −0.49 | 0.13 | −3.76 | .0002 | |
| Medial surprisal $w_{-2}$ | 1.14 | 0.18 | 6.19 | <.0001 | |
| Initial surprisal $w_{-1}$: $I(w_{-1}\|\text{start cue})$ | −0.11 | 0.11 | −0.99 | .321 | |
| Initial surprisal $w_{-2}$: $I(w_{-2}\|\text{start cue})$ | −1.08 | 0.19 | −5.59 | <.0001 | |

*Note.* BIC = Bayesian Information Criterion; LSTM = long short-term memory; FTP = Forward Transitional Probability.

(*Appendix continues*)

**Table A11**
*Final Bayesian Models*

| Predictor | *b* | *EE* | l–95% CI | u–95% CI | rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Model 1: One- versus two-word repetitions | | | | | | | |
| Intercept | −2.49 | 0.24 | −2.98 | −2.05 | 1 | 8,049 | 15,508 |
| Medial surprisal $w_{-1}$ | −0.5 | 0.14 | −0.78 | −0.23 | 1 | 18,148 | 25,861 |
| Medial surprisal $w_{-2}$ | 1.18 | 0.19 | 0.8 | 1.56 | 1 | 12,812 | 22,163 |
| Initial surprisal $w_{-1}$: $I(w_{-1}\|\text{start cue})$ | −0.11 | 0.12 | −0.34 | 0.12 | 1 | 17,728 | 25,174 |
| Initial surprisal $w_{-2}$: $I(w_{-1}\|\text{start cue})$ | −1.12 | 0.2 | −1.53 | −0.73 | 1 | 11,367 | 19,742 |
| Backward surprisal $w_{-2}$: $I(w_{-2}\|w_{-1})$ | −1.58 | 0.17 | −1.92 | −1.26 | 1 | 22,582 | 26,770 |
| Relative surprisal: $I(w_1\|w_{-1}) - I(w_1\|w_{-2}w_{-1})$ | 0.47 | 0.06 | 0.35 | 0.6 | 1 | 31,046 | 28,160 |
| Boundary at $w_{-2}$ | 1.33 | 0.22 | 0.91 | 1.76 | 1 | 17,238 | 25,796 |
| Boundary at $w_{-3}$ | 0.53 | 0.29 | −0.03 | 1.11 | 1 | 21,479 | 28,701 |
| Duration $w_{-2}$ | −0.52 | 0.1 | −0.71 | −0.33 | 1 | 36,894 | 29,450 |
| Model 2: Two- versus three-word repetitions | | | | | | | |
| Intercept | −2.33 | 0.54 | −3.44 | −1.3 | 1 | 9,823 | 17,015 |
| Medial surprisal $w_{-2}$ | −0.94 | 0.32 | −1.62 | −0.34 | 1 | 11,927 | 15,755 |
| Medial surprisal $w_{-3}$ | 1.35 | 0.43 | 0.53 | 2.23 | 1 | 13,533 | 20,901 |
| Initial surprisal $w_{-2}$: $I(w_{-2}\|\text{start cue})$ | 0.83 | 0.33 | 0.2 | 1.51 | 1 | 12,365 | 17,103 |
| Initial surprisal $w_{-3}$: $I(w_{-3}\|\text{start cue})$ | −1.7 | 0.48 | −2.72 | −0.83 | 1 | 10,445 | 16,570 |
| Backward surprisal: $I(w_{-3}\|w_{-2})$ | −1.69 | 0.39 | −2.51 | −0.97 | 1 | 15,706 | 22,290 |
| Relative surprisal: $I(w_1\|w_{-2}w_{-1}) - I(w_1\|w_{-3}w_{-2}w_{-1})$ | 0.26 | 0.13 | 0 | 0.51 | 1 | 31,729 | 30,498 |
| Boundary at $w_{-3}$ | 0.77 | 0.71 | −0.6 | 2.17 | 1 | 20,845 | 26,041 |
| Boundary at $w_{-1}$ | −0.63 | 0.5 | −1.66 | 0.31 | 1 | 15,130 | 20,224 |
| Duration $w_{-3}$ | −0.55 | 0.19 | −0.93 | −0.17 | 1 | 31,870 | 30,236 |

Received September 21, 2017
Revision received April 30, 2021
Accepted May 1, 2021 ■