OPEN MIND

Discoveries in Cognitive Science

an open access journal

Check for updates

Citation: Upadhye, S., & Futrell, R. (2026). Back to the Future: The Role of Past and Future Context Predictability in Incremental Language Production. *Open Mind: Discoveries in Cognitive Science*, *10*, 1745–1780. https://doi.org/10.1162/OPMI.a.397

DOI:
https://doi.org/10.1162/OPMI.a.397

Supplemental Materials:
https://doi.org/10.1162/OPMI.a.397

Received: 10 November 2025
Accepted: 8 September 2026

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Shiva Upadhye
shiva.upadhye@uci.edu

Copyright: © 2026 Shiva Upadhye and Richard Futrell. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The MIT Press

REPORT

Back to the Future: The Role of Past and Future Context Predictability in Incremental Language Production

Shiva Upadhye[1,2] and Richard Futrell[1]

[1]Department of Language Science, University of California, Irvine, Irvine, CA, USA
[2]Department of Psychology, Tufts University, Medford, Massachusetts, MA, USA

Keywords: language production, information-theoretic linguistics, Probabilistic Reduction, lexical planning, language modeling

ABSTRACT

Contextual predictability shapes how we choose and encode words in production. The effects of a word's predictability given preceding or *past* context are generally well-understood in both production and comprehension, but studies of naturalistic production have also revealed a poorly understood yet robust *backward* predictability effect of a word given only its *future* context, which may be linked to future planning. Across two studies of naturalistic speech, we revisit backward predictability using improved operationalizations, introducing **Conditional Pointwise Mutual Information** (conditional PMI), a conceptually motivated information-theoretic measure that quantifies the information shared between a word and the future sequence under the constraints imposed by the past context. Study 1 shows that this measure reproduces predictability effects of future context on phonetic reduction while remaining non-redundant with backward predictability. Study 2 moves from reduction to word choice, examining substitution errors within a novel generative framework that models lexical and contextual influences to predict the identity of the word that surfaces as an error. Within this framework, we find that past-conditioned predictability increases likelihood of erroneous selection, whereas future-conditioned predictability reduces it. In predicting error identity, conditional PMI emerges as the strongest contextual predictor, subsuming backward predictability. Together, these findings illuminate how past and future context shape word choice and encoding, linking contextual predictability to mechanisms of incremental planning in sentence production.

INTRODUCTION

Naturalistic language production shows remarkable variability in *which* words are chosen and *how* they are produced. Much of this variability—particularly in how word-forms are articulated—has been linked to the dynamics of real-time cognitive processing (cf. Bybee & Hopper, 2001; Pierrehumbert et al., 2002; Seyfarth, 2014 for offline representation-based accounts). Under this view, an increase in processing difficulty leads to more robust encoding of a word's phonetic and articulatory detail. Empirically, this relationship is borne out in *probabilistic reduction*: words that are predictable, whether on their own or in context, are more prone to reduction in phonetic detail or word duration (Aylett & Turk, 2004; Balota et al., 1989; Bell et al., 2003, 2009; Hashimoto, 2021; Jaeger & Buz, 2017; Kurumada, 2011; Pimentel et al., 2021; Pluymaekers et al., 2005; Ranjan et al., 2022; Tily et al., 2009; Wolf et al., 2023 *inter alia*).

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

Although the effects of contextual predictability on processing difficulty have been studied extensively in psycholinguistic research, much of this work has focused on **forward predictability**, i.e., the predictability of a word given preceding context. Forward predictability, or its information-theoretic formalization, forward surprisal[1], has received widespread empirical support as an index of incremental processing difficulty in sentence comprehension (Futrell et al., 2020; Hale, 2001; Levy, 2008; Meister et al., 2021; Shain et al., 2024; Wilcox et al., 2020, 2023; Xu et al., 2023) and production (Bell et al., 2009; Dammalapati et al., 2019; Demberg et al., 2012; Harmon & Kapatsinski, 2021; Jaeger, 2006; Jurafsky et al., 2001; Mahowald et al., 2013; Zhan & Levy, 2018).

However, a key asymmetry between online language comprehension and production is that the latter affords access to conceptual and linguistic representations beyond the recently produced sequence. Since sentence production involves transforming an abstract conceptual representation of the speaker's message into a linearized linguistic output, the order in which words are planned need not mirror the order in which they are produced (Bock & Levelt, 1994; Ferreira, 2000; Ferreira et al., 2018; Momma & Ferreira, 2019). Therefore, when producing an utterance incrementally, speakers have access not only to their *past* output, but also, potentially, to planned or desired *future* representations.

Consequently, ease of lexical planning may reflect both the facilitative effect of incremental predictability and the constraining influence of upcoming representations (Figure 2). For example, consider the utterances below:

1. It's not against the law to _____ **alligators across the river**
2. It's not against the law to _____ **alligators through the mail**[2]

The shared preceding context licenses several compatible verbs at position $t = 7$ (e.g., *drink, dream, gamble, jaywalk, dance* etc.) while the following context (**bolded**) in both utterances narrows down word choice to transitive verbs that can take *alligators* as a direct object. Moreover, the following context in (2) imposes a much stronger constraint on word choice: for example, even though the verb *send* can appear in both contexts, it is more predictable given 'through the mail' than 'across the river.' This backward-looking effect of upcoming material has been broadly referred to as **backward predictability**, which captures how predictable a word is under an upcoming or future sequence Figure 1. A number of studies have shown that backward predictability emerges as a robust predictor of planning difficulty in naturalistic speech, as reflected in both acoustic enhancement (Bell et al., 2009; Hashimoto, 2021; Pluymaekers et al., 2005; Ranjan et al., 2022) and the prevalence of disfluencies (Dammalapati et al., 2019, 2021; Goldman-Eisler, 1958; Harmon & Kapatsinski, 2021; Shriberg, 1996), even after controlling for the effects of other probabilistic variables such as lexical frequency and forward predictability.

Prior work has also shown that predictability from the past context (i.e., forward predictability) and future context (i.e., backward predictability) show asymmetric effects on word durations. Notably, Bell et al. (2009) observed that the effects of contextual predictability appeared to be modulated by frequency for function words: while durations of high-frequency function words were sensitive only to predictability from the past, durations of mid-to-low frequency words were affected by predictability from the following word. By contrast, content word

[1] Defined as the negative log probability of the word conditioned on the preceding context, $-\log p(\text{word} \mid \text{context})$ (Shannon, 1948).

[2] Example utterance from Fromkin (2000).

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

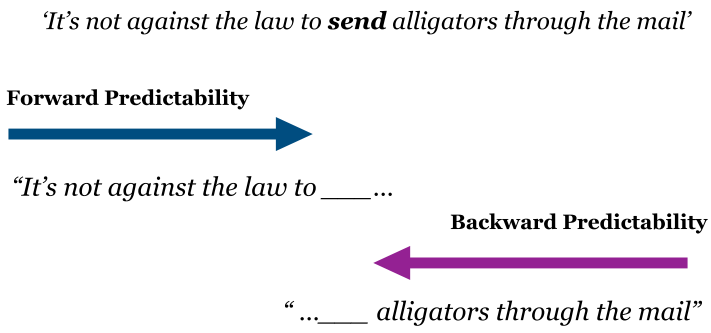


Figure 1. An illustration of forward and backward-looking contextual predictability effects in naturalistic production.

durations were sensitive only to predictability from the following but not the preceding word (cf. Ranjan et al., 2022), potentially reflecting stronger associations between content words and upcoming material.

A number of conceptual interpretations of backward predictability effects have been proposed. In the articulatory-motor domain, Pluymaekers et al. (2005) propose a link between the backward predictability effect and anticipatory motor planning: when the current and following word are highly informative about each other, speakers speed-through the articulation of the current word to initiate motor planning of the upcoming word, resulting in reduced durations. This account is *prospective* in that it attributes (non-)reduction to the availability of upcoming material. In contrast, Bell et al. (2009) put forth a *retrospective* explanation of this effect: predictability—whether unigram, forward or backward—facilitates retrieval of the current word, leading to reduced buffering and smoother coordination between planning and articulation. Another interpretation of this quantity comes from Harmon and Kapatsinski (2021), who treat the predictability of the preceding sequence conditioned on the present word as form of *reactivation*. Once reactivated (through repetition), the preceding sequence serves as a cue that facilitates retrieval of the upcoming word.

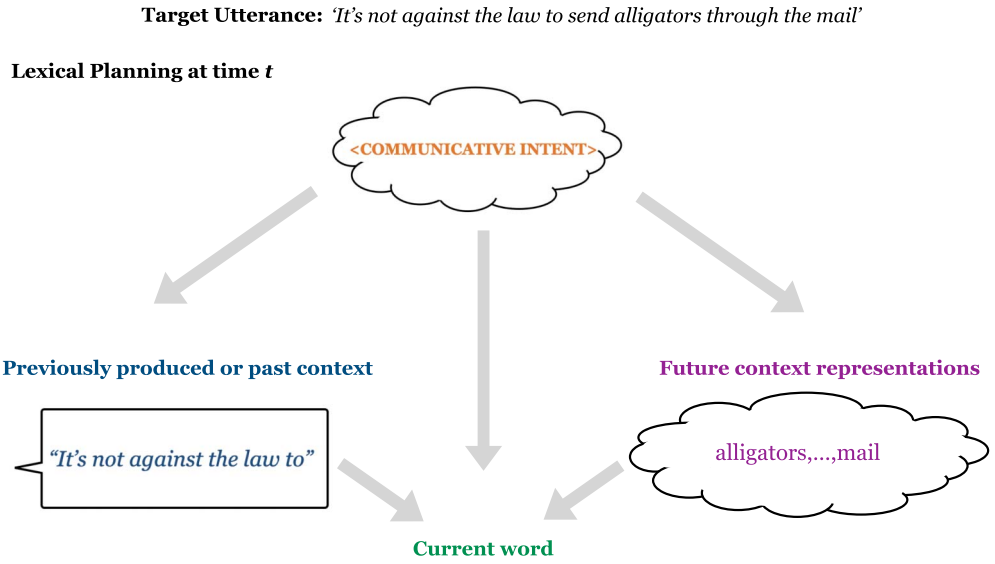


Figure 2. Effects of communicative intent and context-based information sources on lexical planning in incremental language production.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

One barrier to finding a clear interpretation of the backward predictability effect is that we do not yet have the full empirical picture of the effects of predictability on word duration and choice. A key limitation of previous work is that the commonly-used measure of backward predictability assumes implicitly that the future context is produced *independently* of the previously produced or past context. That is, although the measure captures the predictive effects of future material on the present word, it disregards the specific context in which the speaker produces that future material.

Crucially, such an assumption is inconsistent with our understanding of the architecture underlying sentence production. Incremental production begins with conceptual, order-agnostic representations of the speaker's message, which are subsequently *serialized* to satisfy both typological and articulatory-motor constraints (Bock & Levelt, 1994; Lashley et al., 1951; Levelt, 1981). Therefore, the *past* and *future* relative to a notional *present* word may not be intrinsic to how speakers plan their utterance; rather, they reflect a property of how the overt production of the observed utterance unfolds over time. Since both are jointly constrained by the speaker's message, past and future context can be viewed as observable approximations of it.

Moreover, cross-linguistic evidence shows that production planning can proceed *non-linearly*, driven by structural relations between words that may be positioned arbitrarily far apart in the serialized form of the utterance (Lee et al., 2013; Momma & Ferreira, 2019; Momma et al., 2016; Sauppe, 2017; Schriefers et al., 1998). These findings suggest that previously produced and planned yet unrealized future words may be directly linked via relations that are not mediated by the choice of the current word.

However, since backward predictability measures only the predictability of the current word from the future context and ignores the past context altogether, it cannot capture such direct associations between words in the past and future. Beyond this conceptual limitation, the interpretation of backward predictability in regression models that also include forward predictability is complicated by the fact that the two probabilistic variables are highly collinear (Bell et al., 2009), which obfuscates the distinct predictive contributions of past and future context.

Here, we give a more complete characterization of the predictive effects of past and future context on speaker choices in naturalistic production. Conceptually, we argue that backward predictability reflects one particular operationalization of future-context predictability—one that assumes independent effects of the past and future contexts. We propose an alternate mathematical operationalization of future context effects based on the **Conditional Pointwise Mutual Information (PMI)** (Fano & Hawkins, 1961) of the current word and future after extracting information from the past context—a measure that avoids the independence assumption implicit in backward predictability. Methodologically, we address concerns about collinearity and language modeling to improve the estimation of contextual predictability measures.

Empirically, this work is grounded in two studies that illuminate the nuanced effects of contextual predictability on lexical planning. Our first study revisits *probabilistic reduction* as a testbed for evaluating the improved contextual measures, including our proposed conditional PMI-based formulation of future context predictability. Our second study investigates the effects of context on word choice by independently modeling how lexical availability, communicative alignment, and past- and future-context predictability shape the content of lexical substitutions in naturalistic productions.

Our findings have implications for accounts of probabilistic reduction and incremental planning in sentence production. Through our first study, we demonstrate that our principled conditional PMI-based measure qualitatively replicates the effects of backward predictability

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

in predicting word durations while also explaining non-redundant variance in duration. However, it does not reproduce the asymmetric effects of context on function and content words observed by Bell et al. (2009). Instead, we find that lexical and contextual predictability from either direction predict reduction for both content and function words, although function word durations exhibit reduced sensitivity to predictability overall. Finally, our model of substitution errors broadens the empirical scope of research on trade-offs in word choice, which has so far been examined largely independently of the informational constraints imposed by sentential context. This analysis reveals distinct functional pressures imposed by past and future context on the choice of substitution errors, corresponding respectively to goal-invariant (i.e., availability-based) and goal-directed mechanisms in sentence planning.

PROPOSED FRAMEWORK

The goal of the present study is to estimate the distinct contributions of contextual predictability from the past and future on articulatory duration and word choice in naturalistic production. For the speaker, when future representations are entirely unavailable, the current word may be predicted solely from its past context (Figure 3a), as quantified by forward predictability:

$$p(w_t|C_{<t}) \tag{1}$$

where w_t is a word at time t and $C_{<t} = w_1 \cdots w_{t-1}$ is the previously produced or past context. Backward predictability provides an analog of this effect in the reversed direction:

$$p(w_t|C_{>t}) \tag{2}$$

where $C_{>t} = w_{t+1} \cdots w_N$ is an upcoming sequence of words, with w_N serving as the last word of the sentence. This measure can be seen as quantifying the predictability of the current word

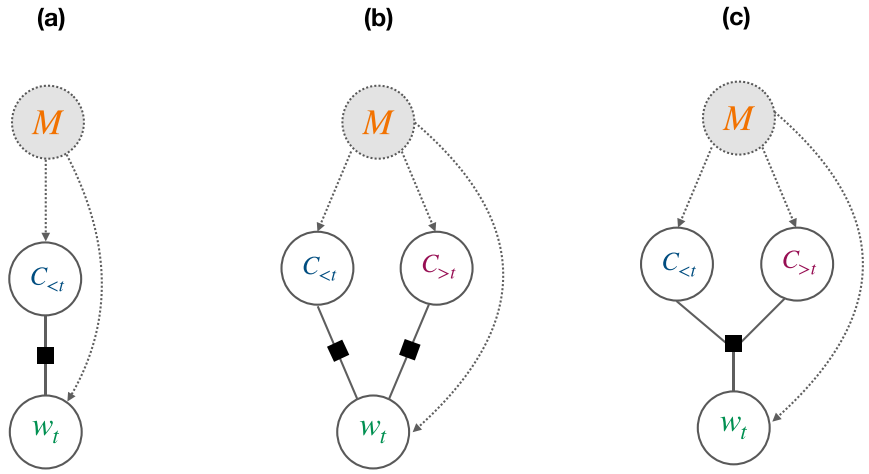


Figure 3. An illustration of the information-processing dependencies between the speaker's message (M), past context ($C_{<t}$), future context ($C_{>t}$), and the current word (w_t). Here, we treat M as a latent variable, and $C_{>t}$ as an observed variable, even though for the speaker, the future is not realized until after the production of w_t. Solid lines indicate explicit conditioning dependencies while dashed lines indicate causal influences between the latent variable M and the contextual representations. Black squares indicate **factors**, which define functions between connected variables. For example, the black square in (c) denotes that w_t is a function of both $C_{<t}$ and $C_{>t}$, whereas in (b), w_t is determined by independent functions of $C_{<t}$ and $C_{>t}$. Left (a): Current word depends solely on past context, as reflected in forward predictability. Middle (b): the current word is influenced separately by past and future context—the assumption implicit in backward predictability. Right (c): Current word is jointly influenced by the past and future context, as in the case of bidirectional word probability.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

under a future sequence, for a speaker who has more or less fully planned the future context as it will be produced.

As in some previous work (e.g., Pluymaekers et al., 2005), backward associations between the current word and the future word or sequence can also be modeled using (unconditional) Pointwise Mutual Information or PMI:

$$\text{PMI}(w_t; C_{>t}) = \log p(w_t|C_{>t}) - \log p(w_t) \tag{3}$$

where $\text{PMI}(w_t; C_{>t})$ measures the strength of association between the word w_t and the future $C_{>t}$ after accounting for how frequently they occur independently. Thus, a positive value of $\text{PMI}(w_t; C_{>t})$ indicates that the word and the future sequence are highly informative about each other i.e., are more likely to co-occur than would be predicted by chance alone.

However, when viewed as an index of future planning difficulty, backward predictability, or its information-theoretic analog $\text{PMI}(w_t; C_{>t})$, is limited in one crucial respect: it treats the speaker's current state—i.e., the context produced so far—as irrelevant to upcoming production. Formally, it models the conditional dependence of the current word (w_t) on the future context ($C_{>t}$), but disregards the (past) context ($C_{<t}$) in which both are anchored. In effect, this treats past and future context as independent sources of information about w_t (Figure 3b)—an assumption that does not hold, since both are ultimately shaped by the speaker's intended message (M). Instead, if the past and future jointly constrain the current word (Figure 3c), then predictability from the future alone may over- or underestimate how *uniquely* informative that future context is.

Quantifying Future Context Predictability

In this work, we relax the assumption that past and future context have independent predictive effects on lexical planning. With this goal in mind, we propose a principled alternative to backward predictability, based on the **conditional PMI** of the current word w_t and the future context $C_{>t}$ given the past context $C_{<t}$:

$$\text{conditional PMI}(w_t;\ C_{>t}|C_{<t}) = \log \frac{p(w_t|C_{>t}, C_{<t})}{p(w_t|C_{<t})} \tag{4}$$

$$= \underbrace{\log p(w_t|C_{>t}, C_{<t})}_{\text{Bidirectional word probability}} - \underbrace{\log p(w_t|C_{<t})}_{\text{Forward word probability}} \tag{5}$$

The bidirectional probability term in Equation 5 quantifies how past and future context *jointly* predict the current word, thereby preserving any associations between non-adjacent words within the utterance. Subtracting the forward probability then isolates how much *more informative* the future context is about the current word, given the speaker's current state, i.e., the context they have already produced. It is this second step that distinguishes our measure from prior work that uses either vanilla backward probability or bidirectional probability, neither of which isolates the unique contribution of future context predictability once the past has been taken into account. Wolf et al. (2023), for instance, show that bidirectional context shares higher mutual information with word duration than past context alone; however, because their measure does not extract out the past, it cannot establish how much of that gain is attributable to future context specifically.

In comparison, backward predictability provides an estimate of future context predictability that does not take into account the already produced context and can therefore be considered *context-independent* across all possible past sequences that the speaker could have

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

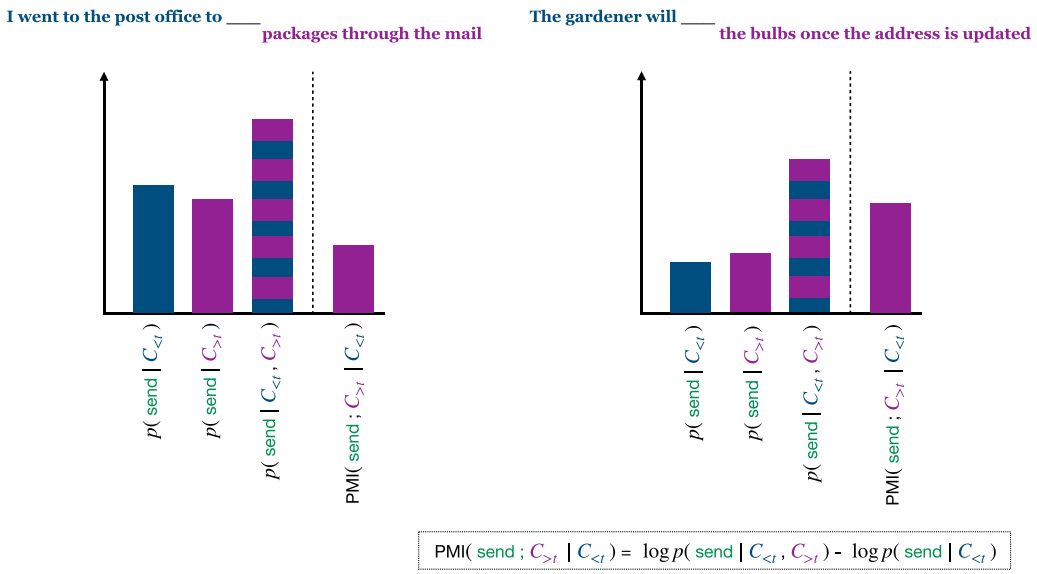


Figure 4. Simulated comparison of forward probability $p(w_t|C_{<t})$ (in blue), backward probability $p(w_t|C_{>t})$ (in purple), bidirectional probability $p(w_t|C_{<t}, C_{>t})$ (striped blue and purple), and conditional **PMI**$(w_t ; C_{>t} | C_{<t})$ (in purple) to illustrate when backward predictability and conditional PMI diverge in their estimations of the predictive contribution of future context, with a focus on how backward predictability may over- or underestimate this contribution. Recall that conditional PMI is defined as the difference between log-transformed bidirectional and forward probabilities. The y-axis is illustrative and reflects only the relative magnitude of each quantity; it should be noted, however, that the forward, backward, and bidirectional probabilities are probabilities, whereas conditional PMI is in nats.

produced[3]. Cognitively, it corresponds to a state of epistemic uncertainty in which the speaker aggregates over all possible past sequences: that is, a state where the realized past itself is treated as no more privileged than the alternatives.

A positive value of conditional PMI suggests that the current word and future sequence are informative about each other *beyond* what can be predicted by the past context. For instance, consider the following sentences:

3. It's not against the law to **[send]** alligators through the mail
4. She poured the **[gin]** into her laptop

Ex. (3) provides an example of positive conditional PMI: the current word (*send*) becomes more predictable in bidirectional context, which includes the informative upcoming sequence *through the mail*. However, (4) illustrates an instance where conditional PMI may be negative, as the current word (*gin*) is highly predictable under the past context, but becomes less predictable once the future is taken into account.

Moreover, as noted earlier, by explicitly controlling for the contribution of the past context, conditional PMI more accurately captures the unique information gained from the future context (Figure 4). For example, consider the sentence below:

5. I went to the post office to **[send]** packages through the mail

Unlike (3), where the past context is not predictive of the target *send*, (5) illustrates an instance where the preceding sequence "I went to the post office to" strongly predicts *send*

[3] Put differently, backward predictability can be viewed as analogous to the bidirectional probability in Equation 5 but with the past context marginalized out i.e., $\sum_c p(w|C_{>t}, C_{<t} = c) \cdot p(C_{<t} = c)$.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

such that the *additional* information gained from the future context is limited. Because backward predictability quantifies the predictive contribution of the future independent of the past context, it may, in this case, overestimate how uniquely informative the future context "packages through the mail" is about the target. As noted earlier, conditional PMI is defined as the difference between log-transformed bidirectional and forward probabilities; consequently, when both of these probabilities are high, their difference is low, reflecting a limited predictive contribution of the future context. Crucially, this cannot be captured by backward predictability, whose value depends solely on whether the predictability from the future context is high.

Similarly, backward predictability may also underestimate the predictive influence of the future context. Consider the example below:

6. The gardener will **[send]** the bulbs as soon as the customer's address is updated.

In (6), the past context does not strongly predict the target *send* and instead favors verbs such as *plant, grow*, or *prune*. Given future context alone, *send* or *ship* may become moderately predictable, such that the backward predictability of *send* may exceed its forward predictability. Yet, the future context is ambiguous in isolation: since *bulbs* also admits the *electric bulb* sense of the word, the future also licenses verbs such as *repair, change*, or *install*. This ambiguous reading of the future context is resolved when considered in combination with the past context. Although the past is less constraining on its own, it provides the disambiguating cue *gardener*, which rules out candidates such as *repair* or *install*, while verbs such as *send* or *ship* still remain more compatible with the future context. The future is therefore more informative about the target when anchored in the past than backward predictability reveals in this particular example.

Beyond this, a key mathematical property of Pointwise Mutual Information is that it is symmetric: it does not commit to whether the future sequence influences the choice of the current word or vice versa. A positive value of conditional PMI is, therefore, compatible with (i) a *retrospective* interpretation, i.e., the planned future sequence is highly informative about word choice at the current time-step, or (ii) a *prospective* interpretation, wherein the choice of the current word facilitates production of upcoming material.

This latter interpretation is broadly aligned with the concept of *value-to-go* (Sutton & Barto, 1998; Todorov, 2009) in action planning, optimal control, and reinforcement learning. Value-to-go is the total expected value of taking an action, consisting of the immediate utility of the action itself and the expected value of future actions after it. In theories of optimal planning, an agent chooses actions to maximize this value-to-go function. As applied to language production, value-to-go includes predictability as a factor—more predictable actions or words are higher value because they incur less cognitive cost, and can be produced in a more routinized way (Futrell, 2023; Gershman, 2020; Gershman & Bhui, 2020; Lai & Gershman, 2021). Therefore, an optimal planning agent will tend to produce a word w_t that increases the predictability of a high-value *future* sequence $C_{>t}$, as measured by the surprisal

$$-\log p(C_{>t}|w_t, C_{<t}). \tag{6}$$

We show how conditional PMI can be derived from Equation 6 in Appendix B.

To enable a principled comparison between conditional PMI and backward predictability, we must account for the fact that backward predictability is itself correlated with forward predictability, whereas conditional PMI partials out the standalone contribution of the past,

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

effectively producing a decorrelated quantity. We therefore evaluate conditional PMI against the following decorrelated alternative to backward predictability[4]:

$$\text{Relative Backward Predictability} = \underbrace{\log p(w_t|C_{>t})}_{\text{Backward Conditional Probability}} - \underbrace{\log p(w_t|C_{<t})}_{\text{Forward Conditional Probability}} \quad (7)$$

Both relative backward predictability and conditional PMI partial out the standalone predictive contribution of the past, but differ in that conditional PMI preserves the joint effects of past and future, whereas relative backward predictability treats past and future as independent sources of predictability. From an information-theoretic perspective, relative backward predictability can be viewed as a log-likelihood ratio that represents how much *more informative* the upcoming sequence is about the current word *compared* to the past context. A positive value of relative backward predictability reflects that the future context outperforms the past in terms of predicting the current word while a negative value reflects the opposite.

It should be noted that relative backward predictability can be viewed as an *asymmetric* formulation of future context predictability since it reflects the assumption that it is the future context that influences the current word. By contrast, conditional PMI is *symmetric*: it measures the association between the current word and future sequence given past context, but is agnostic about the direction of influence between the current word and the future context. For a full comparison, we can also consider the unconditional PMI in Equation 3, which is an alternative symmetric formulation of future context predictability that, like relative backward predictability, assumes independent effects of past and future. We discuss effects of unconditional PMI relative to other formulations of future context predictability in Appendix E.1.

Language Modeling

In addition to the theoretical and modeling considerations addressed above, the estimation of forward predictability, backward predictability, and PMI presents a non-trivial challenge. Much of the prior work on probabilistic reduction focused primarily on bigram contexts, where both forward and backward predictability could be estimated from the joint frequency of the bigram sequence and the respective unigram frequencies of the individual words (Bell et al., 2009; Pluymaekers et al., 2005). More recently, among studies of speech that have investigated the effects of longer contexts on reduction, it has been common practice to train separate *neural* language models (LMs) for forward (left-to-right) ($\overrightarrow{p}$) and backward (right-to-left) ($\overleftarrow{p}$) predictability, with the backward predictability model being trained on a reversed corpus (Dammalapati et al., 2021; Harmon & Kapatsinski, 2021; Ranjan et al., 2022).

However, this practice is not suitable for estimating our measures. First, estimating conditional PMI involves estimating the bidirectional or *infill* probability ($\overleftrightarrow{p}$), which cannot be feasibly estimated from a next-word prediction-based or autoregressive language model.[5] Second, if we

[4] A commonly used alternative method for de-correlating predictors is **residualization**—a process that involves regressing the variable of interest on a correlated variable and using the residuals or prediction errors in lieu of the target variable. While residualization can effectively de-correlate predictors by removing covariance between these variables, this process may alter the interpretability of the construct represented by the target variable (Breaugh, 2006; Wurm & Fisicaro, 2014). In other words, a residualized variant of backward predictability may address the problem of collinearity, but may render this effect less cognitively interpretable. See also (Opedal et al., 2024) for critiques of residualization in the context of predictability effects on reading times.

[5] Existing pre-trained LMs capable of estimating masked bidirectional probabilities, such as BERT (Devlin et al., 2018), are not suitable for this task, because they not only differ from autoregressive models in terms of training data, but also in the training objective and model architecture. Furthermore, masked LMs such as BERT are not generative models and do not define a joint distribution over words.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

measure forward and backward predictability using separately-trained models, we introduce a potential confound, because the probability distribution modeled by a neural LM depends on both the nature of the neural architecture itself and the variability introduced in the training process (for example, the contents of the random mini-batches used in training by stochastic gradient descent, and the random initialization of parameters). If one trains separate forward and backward LMs, their outputs may not only reflect differences in the training input, but also additional sources of variance resulting from the training process (Fehlauer et al., 2025).

For these reasons, a principled comparison between the contextual predictability variables requires that forward, backward, and infill probabilities all be derived from a *single* model $\overrightarrow{p}_\theta$. To train such a model, we adapt prior work on enabling infill probability estimation from autoregressive LMs (Bavarian et al., 2022; Donahue et al., 2020). In principle, this involves selecting a word at random within the utterance, and re-arranging the positions of the word, its preceding context, and following context as follows (see also Figure 5):

7. Original utterance: So this is the **first** time I did this conversation
8. Augmented training input: <PRE> So this is the <SUF> time I did this conversation <MID> **first** <eos>

Here, the selected word (**first**) is moved to the end of the utterance such that the model learns to predict this word given both the past and future context that appear *before* it, demarcated using <PRE> and <SUF> tokens, respectively. This allows us to use an autoregressive LM trained on next-word prediction while supplying it with bidirectional context. Furthermore, previous work has shown that training or fine-tuning LMs on naturalistic dialogue improves

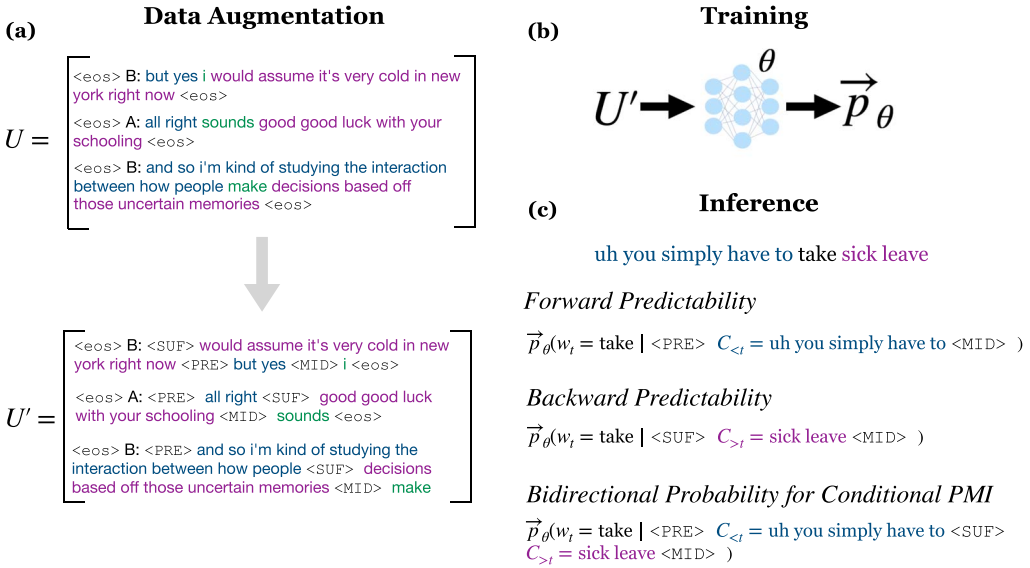


Figure 5. Overview of the process for estimating contextual predictability variables from a custom-trained language model (LM). (a) Data augmentation process for enabling estimation of all three probabilities from an autoregressive LM. U is the original corpus of utterances while U' is the augmented corpus. Each utterance (u) in U was transformed by uniformly sampling a position in the utterance, selecting the word in that position, and appending this word to the end of the utterance. The past and future context with respect to the original word position are demarcated using <PRE> and <SUF> tokens, and the transposed word is preceded by a <MID> token. In 50% of the utterances, the positions of the <SUF> and <PRE> contexts were swapped. This was done in accordance with prior work, which has found that changing the order of the preceding and following sequences improves estimation of infill probabilities (Bavarian et al., 2022). See Appendix C for an algorithmic implementation of this process. (b) U' serves as the training input to a randomly initiated GPT-2 language model parameterized by θ. (c) An illustration of the inference process for estimating forward, backward, and bidirectional probabilities from the trained GPT-2 model $\overrightarrow{p}_\theta$.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

the model's sensitivity to speech-related behaviors compared to those trained on written text alone (Umair et al., 2024). Since our study focuses on signatures of production difficulty in naturalistic conversations, we train a GPT-2 small (124 million parameter) transformer language model (Radford et al., 2019) and a word-level tokenizer on the CANDOR corpus of spontaneous speech (Reece et al., 2023). Additionally, we also prepend each utterance with a speaker tag (A or B to indicate a change in turn) to account for partner-tracking, which has been shown to affect predictability estimates in conversational speech (Warnke & de Ruiter, 2023). Model and training-related hyperparameters were evaluated through grid search, and the model that yielded the lowest perplexity on the Switchboard corpus was selected for deriving the probabilistic measures (see Appendix C for training details). The inference process for estimating forward predictability, backward predictability, and conditional PMI from this model ($\overrightarrow{p_\theta}$) is detailed in Figure 5(c). Log-transformed forward and backward probabilities were plugged into Equation 7 to compute relative backward predictability. Likewise, conditional PMI was computed by plugging in log-transformed bidirectional and forward probabilities into Equation 5. Further details of the data augmentation and model training process can be found in Appendix C. All the contextual predictability measures used in both our studies were estimated using this language modeling approach.

To evaluate how closely forward and backward probabilities obtained from this training approach aligned with those estimated from separate models, we also trained separate forward ($\overrightarrow{p_\psi}$) and backward ($\overleftarrow{p_\varphi}$) GPT-2 models on the unmodified CANDOR corpus. Forward probabilities estimated from $\overrightarrow{p_\theta}$ were highly correlated with those obtained from $\overrightarrow{p_\psi}$ ($r = 0.85$, $p < 0.001$), and a similar correlation was observed with backward probabilities obtained from the infill-trained ($\overrightarrow{p_\theta}$) and backward-trained ($\overleftarrow{p_\varphi}$) models ($r = 0.83$, $p < 0.001$). This suggests that our infill-trained model provides reliable estimates of the predictability values derived from separate models.

The modeling paradigm described in this section improves upon alternatives involving separate models in that it enables computationally tractable estimation of forward, backward, and infill probabilities from the same underlying distribution $\overrightarrow{p_\theta}$. We want to emphasize that we do not consider the infill-augmented GPT-2 as instantiating a particular cognitive claim about how speakers estimate probabilities during on-line production. Rather, we view the various probabilistic quantities estimated from this model as approximating the kinds of contextual information that may be available to speakers during incremental production, which may be arrived at by many different means.

We choose to use the GPT-2 architecture for estimating probabilities because this provides the most accurate estimate of all the information in the context that might be available to a producer. An open question in the literature concerns the amount of context appropriate for explaining phonetic reduction, and recent work by Jacobs et al. (2025) suggests that highly local *n*-gram–based estimates of forward and backward predictability may be sufficient to account for reduction, when combined with information about word position or prosodic structure.

Though interesting, we primarily leave questions about the explanatory adequacy of different context window sizes to future work, although we also report analyses with *n*-gram predictors in Appendix E.2.

STUDY 1: REVISITING PREDICTABILITY EFFECTS ON WORD DURATIONS

Previous work on backward predictability effects in production has largely focused on modeling disfluency (Bell et al., 2003; Dammalapati et al., 2019, 2021; Harmon & Kapatsinski,

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

2021) or articulatory realization in speech (Bell et al., 2009; Hashimoto, 2021; Ranjan et al., 2022). Among studies of articulatory reduction, a robust finding is that backward predictability often emerges as a stronger contextual predictor of reduction than forward predictability in both spontaneous and read-aloud speech (Bell et al., 2009; Hashimoto, 2021; Ranjan et al., 2022). A notable result reported by Bell et al. (2009) is that bigram backward predictability strongly correlates with reduced durations in content and mid-to-low-frequency words, but does not significantly modulate reduction in high-frequency function words, which are instead highly sensitive to bigram forward predictability. By comparison, Hashimoto (2021) finds that predictability from the preceding morpheme also affects the articulatory realization of content morphemes in conversational Japanese speech, although backward predictability remains stronger overall. Furthermore, Ranjan et al. (2022) observe that the asymmetric effects of forward and backward predictability across lexical classes reported by Bell et al. are not borne out in the durations of function versus content words in read-aloud Hindi speech.

In this study, we aim to revisit the analysis of Bell et al. (2009) with two objectives in mind. First, we seek to conduct a controlled comparison of relative backward predictability and conditional PMI to examine how the assumption of (in)dependence between the past and future context affects the magnitude of the future context predictability effect. Because prior work has revealed inconsistent effects of forward and backward predictability on the realization of function versus content words, a secondary aim of this study is to re-examine the effects of contextual predictability and lexical class on articulatory reduction.

Methods

Materials. Word durations were extracted from Switchboard NXT annotations (Calhoun et al., 2010; Godfrey et al., 1992), which provides word-level alignments and Penn Treebank part of speech (POS) tags (Marcus et al., 1993) along with Tone and Break Indices (ToBI) and disfluency annotations. A key consideration in modeling predictability effects on word duration in naturalistic production concerns the inclusion of disfluencies. Past work has demonstrated that the occurrence of disfluencies correlated with enhanced word duration (Dammalapati et al., 2019), particularly in the case of function words (Bell et al., 2003; Fox Tree & Clark, 1997). Consistent with Bell et al., 2009's analysis, we exclude words in disfluent contexts by limiting our analysis to only include utterances without repair or reparandum annotations, yielding a total of 435,795 words.

Statistical Analysis.

Control Variables. We retain the covariates used by Bell et al. (2009), including word length (in syllables), speaker age, and speaker sex. Two of Bell et al.'s controls are prosodic: speech rate and a binary indicator of whether the word lies at the boundary of an intonational phrase (IP). Speech rate, measured in syllables per second, was computed over the entire utterance. Although Bell et al. (2009) coded both phrase-initial and phrase-final position separately, phrase-initial position did not significantly improve their control model, and they therefore retained only the phrase-final indicator in their core analysis. This asymmetry is consistent with the broader literature, which has documented robust effects of phrase-final lengthening on segmental and word duration (Clopper et al., 2018; Turk & Shattuck-Hufnagel, 2007; Wightman et al., 1992). Because our analysis is intended as a replication of, and comparison against, Bell et al. (2009), we adopt their control variables to preserve interpretability and comparability with their findings. The corpus, however, provides ToBI annotations for only a smaller subset of the data; we therefore exclude the phase-final boundary variable from

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

the analysis reported in this section. See Appendix E.3 for a replication of this analysis with this additional prosodic control.[6]

Probabilistic Variables. Lexical frequency was operationalized as the decontextualized or unigram predictability of a word, which we estimate from a count-based *n*-gram model with Laplace smoothing. To ensure consistency between the language statistics that generate the unigram and contextual predictability variables, the unigram model was also trained on the CANDOR corpus. All of the contextual predictability variables, including forward predictability, relative backward predictability, and conditional PMI, were estimated from the single-model setup described in Language Modeling section.[7]

Regression Methodology. We use linear mixed-effects regression models (Barr et al., 2013) to fit word duration (in milliseconds), with the maximally converging model below serving as the baseline:

Baseline model: Duration ~ log unigram predictability + log forward predictability + word length + speech rate + speaker age + speaker sex + 1 | speaker identity

We then generate two variants of this model with future context predictability: one including relative backward predictability (Model 1a) and another including conditional PMI (Model 1b). Since relative backward predictability is a linear combination of forward and backward predictability (see Equation 7), the explanatory power of this model is exactly equivalent to that of a model with unmodified backward predictability, despite differences in coefficient values (Freedman, 2009).[8] Therefore, we do not consider a variant of the baseline model with unmodified backward predictability. However, see Appendix E.1 for a complete comparison between models with unmodified backward predictability, relative backward predictability, unconditional PMI, and conditional PMI.

We then compare these model variants with the goal of identifying the distinct explanatory contribution of future-context predictability and further characterizing the contributions of relative backward predictability and conditional PMI. First, we conduct pairwise model comparisons between the baseline model and Models 1a and 1b to quantify the additional variance explained by including any measure of future context predictability, and to assess which of the two variants yields a greater improvement in explanatory power. We further compare models 1a and 1b to a model with **both** relative backward predictability and conditional PMI (Model 1c) to examine the extent to which these measures make overlapping versus non-redundant contributions to explaining variance in word durations. Since these regression models are nested, and because the goal of this analysis is to quantify the explanatory power of the

[6] Findings about the explanatory power of relative backward predictability and conditional PMI were qualitatively replicated even after including a binary variable encoding whether the word lies at an IP boundary; however, we note some differences in the effect sizes of the probabilistic predictors, which we discuss in further detail in Appendix E.3.

[7] We also provide replications of Study 1 with bigram predictors in Appendix E.2. Under the modeling assumptions adopted in our main analysis, we do not observe stronger effects of bigram predictors over predictors estimated from GPT-2. However, as Jacobs et al. (2025) point out, these results may be sensitive to modeling choices such as the source of the unigram estimate (SUBTLEXus versus CANDOR), the choice of neural language model (separately trained forward and backward models versus single infill-trained setup), or other covariates such as word position versus speech rate.

[8] This was confirmed empirically by fitting regression models with backward and relative backward predictability as predictors and comparing their goodness of fit.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

Table 1. Regression coefficients from models with relative backward predictability and conditional PMI as operationalizations of future context predictability.

	Model 1a (with relative backward predictability)	Model 1b (with conditional PMI)
(Intercept)	145.981 (1.85)***	129.250 (1.82)***
Unigram Predictability	−23.676 (0.12)***	−24.204 (0.12)***
Past Context Predictability	−3.075 (0.11)***	−1.927 (0.11)***
Future Context Predictability	−32.576 (0.40)***	−37.765 (0.62)***
Word Length (in syllables)	88.569 (0.38)***	88.273 (0.38)***
Speech Rate	−32.767 (0.14)***	−33.163 (0.14)***
Speaker Age	0.070 (0.04)	0.090 (0.104)
Speaker Sex: M	−12.035 (0.86)***	−11.117 (0.84)***

Probabilistic predictors are **bolded**. Parentheses denote standard error. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (*ns*).

two operationalizations of future-context predictability rather than evaluate model complexity, we use the Likelihood Ratio Test to conduct model comparisons.[9]

Finally, to examine the modulatory effects of lexical class on word duration, we set up additional variants of Models 1a–1b that include interactions between lexical class and the probabilistic predictors. Lexical class was coded as a binary factor with FUNCTION as the reference level. The class of each word was determined based on the Penn Treebank POS tags provided in the corpus. Nouns, verbs, adjectives, adverbs, and cardinal numbers were categorized as content words while the rest of the tags (excluding disfluencies and symbols) were categorized as function words[10]. This yielded a total of 416,831 words, with 214,990 content words and 201,841 function words.

See Appendix D.1 for detailed specifications of all the models used in this analysis.

Results

Model Comparison. Regression coefficients from Models 1a and 1b are summarized in Table 1. Positive coefficients predict lengthening of word duration, whereas negative coefficients reflect reduction. In both Models 1a and 1b, word length is positively correlated with duration while faster speech rate is associated with reduction.

As expected, all three probabilistic variables exhibit a significant inverse relationship with word duration (Figure 6a). First, the frequency effect was consistently replicated: higher unigram predictability correlates with reduced word duration in both model variants. Predictability from the past context is likewise negatively correlated with duration, although

[9] We found no difference in the qualitative conclusions obtained from likelihood ratio tests as opposed to a complexity-penalizing model selection criterion such as the Bayesian Information Criterion (BIC). See Appendix G for BIC-based model comparisons.

[10] In certain cases, the same word received different POS tags in the corpus based on context e.g., the word 'really' was tagged as RB (adverb) in 166 instances and UH (disfluency) in 46 instances. In this case, only the occurrences of 'really' that were tagged as 'RB' were mapped to the content class.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

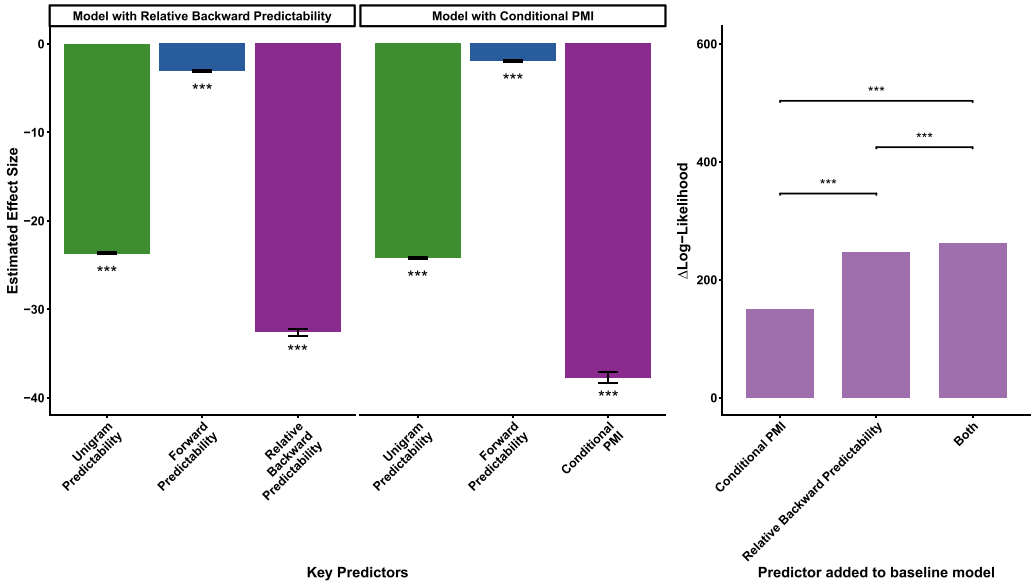


Figure 6. (A) Estimated effect sizes for all probabilistic predictors from models with relative backward predictability and conditional PMI as formulations of future context predictability. Relative backward predictability assumes independence between past and future, whereas conditional PMI assumes conditional dependence. Error bars denote standard error. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (*ns*); (B) Δlog-likelihood values obtained from adding future context predictability measures to the baseline model incrementally. Higher values of **ΔLog-Likelihood** indicate a better fit to the data.

this effect is smaller than that of both unigram predictability and predictability from the future. Among the three probabilistic variables, predictability from the future emerged as the strongest predictor of reduction, with both relative backward ($\beta = -32.576$, $SE = 0.396$, $p < 0.001$) and conditional PMI ($\beta = -37.765$, $SE = 0.618$, $p < 0.001$) exhibiting an inverse effect on duration.

Results of the incremental likelihood ratio tests are presented in Figure 6b. Including predictability from the future improved model fit relative to baseline, both for relative backward predictability (Model 1a: $\Delta\texttt{Loglik} = 3363.199$, $\chi^2 = 6726.5$, $p < 0.001$) and conditional PMI (Model 1b: $\Delta\texttt{LogLik} = 1856.581$, $\chi^2 = 3714$, $p < 0.001$). Moreover, a comparison between models 1a and 1b reveals that relative backward predictability explains more variance in word duration than conditional PMI ($\Delta\texttt{LogLik} = 1506.618$, $\chi^2 = 3012.5$, $p < 0.001$). However, the model with both conditional PMI and relative backward predictability yields a small albeit significant increase in explanatory power compared to the model with relative backward predictability ($\Delta\texttt{LogLik} = 122.166$, $\chi^2 = 245.6$, $p < 0.001$), hence emerging as the strongest performing model of the three.

Effects of Lexical Class and Predictability on Word Duration. All regression coefficients from these analyses are reported in Appendix D.2 and interactions between lexical class and predictability are visualized in Figure 7. Across both Models 1a and 1b variants, there is a significant interaction between unigram predictability and lexical class, with unigram predictability exhibiting a stronger effect on reduction for content words compared to function words. A similar pattern holds for forward predictability, where content word durations show greater attenuation when they are predictable from the past. For function words, the simple effect of forward predictability is small and positive, indicating attenuation and even possible reversal of the reduction effect once other sources of predictability are controlled for. Finally, the regressions reveal significant interactions between lexical class and

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

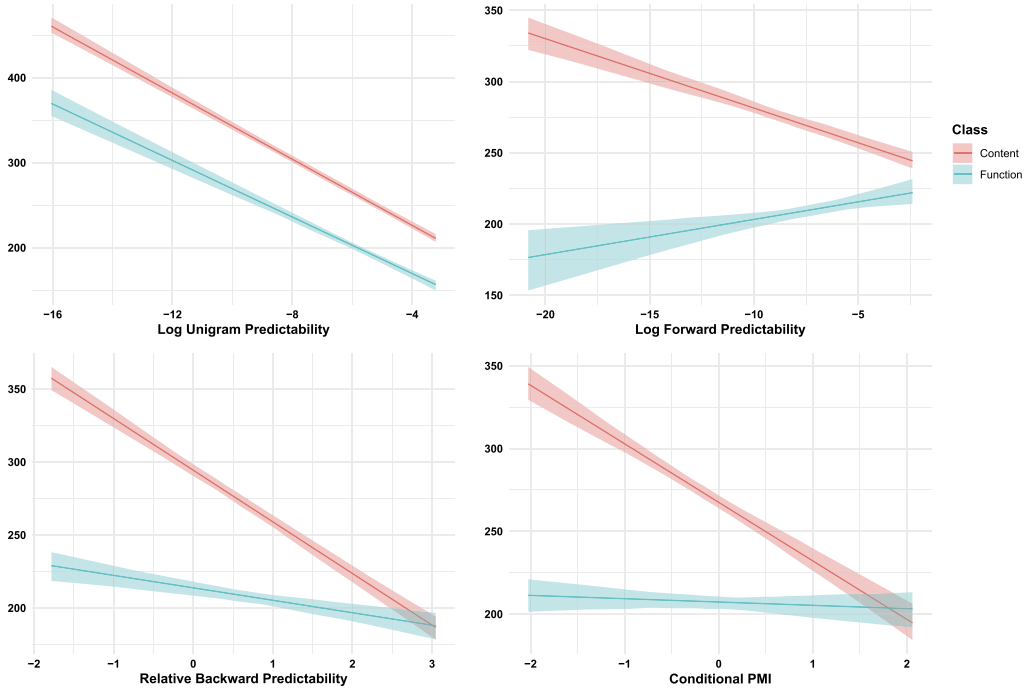


Figure 7. The modulatory effect of lexical category (function/content) on probabilistic reduction, when controlling for other predictors. All predictors are log-transformed. 95% confidence intervals estimated via bootstrapping (N = 1000 simulations). All interactions emerged significant.

future-context predictability, both for relative backward predictability and conditional PMI. Notably, similar to unigram predictability and predictability from the past, content word durations are more strongly modulated by predictability from the future than function word durations.

Discussion

The key findings of this study can be summarized as follows. First, relative backward predictability and conditional PMI yield qualitatively similar effects on reduction, with both measures exhibiting an inverse relationship with word duration. Crucially, predictability from the future, regardless of operationalization, emerges as the strongest probabilistic correlate of reduced durations in our analysis, followed by unigram predictability. By comparison, predictability from the past shows a small albeit robust effect across all model variants.

However, the two operationalizations of future context predictability diverge quantitatively. Our model comparisons reveal that relative backward predictability provides a better fit to word durations than conditional PMI. This pattern is further supported by the observation that, in the model including both measures, relative backward predictability emerges as the strongest contextual predictor of word duration, whereas conditional PMI shows a smaller effect than both relative backward and unigram predictability. It should be noted, however, that backward predictability does not subsume conditional PMI in explaining variance in word duration, as including conditional PMI still yields a significant gain in explanatory power. This suggests that the two operationalizations of future-context predictability are non-redundant variables: conditional PMI contributes complementary information above and beyond what is captured by relative backward predictability. In other words, word duration is influenced

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

to some extent by the predictability of the word given both the past and the future *jointly*, beyond what could be explained by predictability from the past and future separately. We revisit the discussion about these two measures in Study 2's Discussion and General Discussion sections.

A second key finding of this study concerns the function–content asymmetry in wordform variation. Notably, Bell et al. (2009) reported that the effects of forward predictability on articulatory duration were limited to high-frequency function words,[11] with no significant effect on the duration of mid-to-low frequency function and content words. By comparison, backward predictability emerged as the strongest contextual predictor of content and mid-to-low frequency function words. However, our results did not replicate these asymmetric effects of past and future context predictability across lexical classes. Instead, we observe that predictability from both directions predicted reduction in content word durations. The effect of lexical class was, however, borne out in the sensitivity to predictability-induced reduction: content words exhibit stronger predictability-induced reduction across all sources of predictability (unigram, past, and future context), whereas effects for function words are substantially attenuated. Therefore, while both function and content words are sensitive to probabilistic structure regardless of whether predictability is conditioned on previously produced context or unrealized future representations, the degree of sensitivity appears to be modulated by lexical class.

In function words, unigram predictability emerges as the strongest probabilistic correlate of word duration, suggesting that reduction in function words may primarily reflect context-independent lexical accessibility or routinization (Bell et al., 2003; Bybee, 2002; Fox Tree & Clark, 1997; Kapatsinski, 2010). The effects of contextual predictability on the duration of function words is more nuanced. First, we obtain a mild positive relationship between forward predictability and duration of function words when interactions between lexical class, unigram predictability, and future context predictability are included in the model. However, the positive coefficient for forward predictability in function words should be interpreted with caution, as a simpler model without interactions yields the expected negative relationship, suggesting this is likely an artifact of including multiple correlated interaction terms in the model. Comparing the two operationalizations of future context predictability, we find that reduction in function word durations is most strongly driven by relative backward predictability while sensitivity to conditional PMI is attenuated. We discuss the conceptual implications of these results for the two operationalizations of future context predictability in General Discussion section.

It bears mentioning that our analysis of the function–content distinction in probabilistic reduction differs from Bell et al. (2009) along two dimensions: the size of the context window and the language models used for estimating predictability. First, Bell et al. (2009) restricted contextual windows to the immediately preceding or following word and estimated forward and backward predictability using count-based bigram models. In comparison, our study employed the GPT-2 language model, which conditions on broader context and produces smoothed predictability estimates derived from a continuous representational space (Kapatsinski, 2026). Other recent studies using neural language models (e.g., LSTMs or transformers) have likewise reported significant effects of forward predictability or surprisal on content word durations (Clark et al., 2025; Ranjan et al., 2022). One possible explanation for this divergence is that these results may be sensitive to the differences in predictability estimates derived from discrete *n*-gram vs continuous neural language models.

[11] These include ten of the most high frequency function words i.e., *a, the, in, of, to, and, I, it, you*.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

STUDY 2: MODELING SUBSTITUTION ERRORS IN NATURALISTIC PRODUCTIONS

The above study presents a controlled comparison between two alternative operationalizations of future context predictability, both of which exhibit qualitatively similar behavior to backward predictability in modulating articulatory duration. While we did not observe differential effects of past and future context predictability in the reduction of function versus content words, word durations were nonetheless more sensitive to predictability from the future than from the past. This observation is consistent with broader findings that backward predictability emerges as the strongest contextual predictor of other planning difficulties such as filled pauses and repetitions (Dammalapati et al., 2019, 2021; Harmon & Kapatsinski, 2021; Shriberg, 1996).

However, modeling reduction or disfluency may not be the most suitable paradigm for investigating the distinct influences of past and future context on lexical planning. These phenomena primarily index the ease of cognitive processing during production, with unpredictability from either direction correlating with an increase in planning difficulty. Therefore, they offer a limited window into the sentence planning mechanisms by which context influences speaker choices during lexical planning. For example, consider the following utterances from the Switchboard corpus:

9. Well what happens is that if people start becoming having chronic illness and and things like that
10. Uh you simply have to take accumulate your sick leave. And take your sick leave

Ex. (9) and (10) illustrate the distinct pressures imposed by past and future context on word choice. Whereas the preceding context may steer the speaker into selecting *becoming* in (9), this choice is incompatible with the upcoming material, which strongly favors *having*. Similarly, *take* is a more frequent and predictable choice compared to *accumulate* in (10), despite being incompatible with both the future context and the speaker's message. In both cases, incompatibility with future plans and intended semantics incentivizes the speaker to initiate a repair.

Prior work has shown that lexical accessibility and semantic alignment can similarly exert competing influences on word choice, leading to mis-selections or *good-enough choices* (Ferreira & Griffin, 2003; Goldberg & Ferreira, 2022; Koranda et al., 2022; Rapp & Samuel, 2002). Because speakers operate under information processing constraints, lexical selection may reflect a trade-off between automatic or **goal-invariant** strands of processing that prioritize *economy of effort* and controlled or **goal-directed** processing that aims to satisfy the speaker's communicative intent (Bock, 1982; Ferreira & Pashler, 2002; Futrell, 2023; Hartsuiker & Moors, 2017). Through independent manipulation of availability and alignment, Koranda et al. (2022) demonstrated empirical evidence of this trade-off in a gamified word production paradigm: speakers often preferred highly frequent but semantically imprecise alternatives over infrequent yet semantically precise choices to describe a production target.

Beyond word production, previous work has shown that highly incremental (i.e., left-to-right) planning is heavily influenced by conceptual and lexical accessibility (Bock & Irwin, 1980; Dell et al., 2008; Gleitman et al., 2007; Griffin, 2001; Iwasaki, 2011; McDonald et al., 1993; Momma & Ferreira, 2019). This preference for efficiency is reflected in the principle of immediate mention (Ferreira & Dell, 2000): speakers choose to prioritize the overt production of highly available lexical representations to routinely free-up working memory resources (Christiansen & Chater, 2016; Ferreira & Dell, 2000; Ferreira & Swets, 2002; Slevc, 2011). Under this accessibility-driven strategy, the previously produced context may steer the speaker toward highly predictable lexical choices, potentially at the cost of alignment with desired

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

semantics or upcoming context. In contrast, advance planning of the upcoming sequence may impose greater cognitive demands on the speaker since lexical representations need to be planned in advance and sustained in working memory for much longer (Lee et al., 2013; Momma & Ferreira, 2019; Wagner et al., 2010).

Therefore, one possible interpretation that we seek to explore in this study is that past and future context exert competing influences of word choice, reflecting aspects of goal-invariant and goal-directed aspects of processing, respectively. Previously, the study of trade-offs in word choice has been restricted to single-word production since it enables precise manipulation of availability and alignment, but neglects the role of context. In this study, we focus on modeling word choice in naturalistic substitution contexts such as (9)–(10), where the observed self-repair (e.g., having in Ex. 9) provides a reasonable approximation of the speaker's intent or production target. While previous studies have modeled the effects of frequency on the choice of substitution errors (Kapatsinski, 2010) and the effects of contextual predictability on the *occurrence* of substitution disfluencies (Dammalapati et al., 2019), our proposed paradigm aims to predict the *content* or identity of the substitution error using a set of theoretically-motivated predictors that index lexical accessibility, semantic and form-based alignment with the intended word, and context-based mechanisms of sentence production.

Methods

Materials. Utterances with naturally-occurring lexical substitutions and self-repairs were identified and extracted from Switchboard NXT annotations using three criteria. First, utterances with an unequal number of *reparandum* and *repair* words were excluded to avoid instances where the speaker may have revised the structural plan of the sentence. Second, we consider two types of repair: (i) where the self-repair immediately followed the reparandum (Ex. 9) and (ii) where the self-repair was preceded by a repetition disfluency or filled pause (see Ex. 11 below).

11. So until I see the entire quote old guard of the soviet military of the soviet government completely roll over and disappear preferably buried, I still consider them a threat

Finally, we apply a grammatical category constraint, which restricts our set of utterances to those where the reparandum and repair had the same Penn Treebank part of speech category. The selected utterances were then processed into utterance frames consisting of the context preceding the error ($C_{<t}$) and the context following the error ($C_{>t}$) with intervening disfluencies removed. Utterances that contained N substitution errors were processed into N distinct frames, with each frame containing one substitution position (see Appendix I for an example). This process yielded a total of 773 substitution utterances.

Model of Substitution Errors. We propose a framework for modeling substitution errors that captures the distinct influences of lexical frequency, contextual predictability, and communicative utility on lexical mis-selection. Within this framework, we aim to predict which word is produced at the substitution position t in the utterance context, framing this task as a logistic regression model that treats the observed error as the positive class and all other alternatives as negative instances. In practice, each word in the vocabulary is treated as a candidate for production at the substitution position, and the model estimates which properties of candidate words increase the probability that a given word is produced as the substitution error. Therefore, predictors in this model indicate whether a given property facilitates or inhibits lexical mis-selection.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

In addition to lexical frequency and contextual predictability, we also consider the **communicative utility or reward** of choosing a given word, defined as its proximity to the self-repair or intended target w_t^* along both semantic and phonological dimensions. Therefore, variables in this model are functions of a possible word w_t, the entire utterance context C, and the intended target w_t^*.

Operationalizing Communicative Reward. Although the observed self-repair provides an approximation of the speaker's communicative intent, several lexical representations may be compatible with the intended semantics. Because semantic and phonological processes interact and jointly constrain lexical selection within the production system, speakers may trade off semantic alignment for lexical accessibility, particularly when multiple representations approximate the intended lexical semantics (Cutting & Ferreira, 1999; Dell & O'Seaghdha, 1992; Ferreira & Griffin, 2003; Goldrick et al., 2011; Koranda et al., 2022; Rapp & Goldrick, 2000). Even when speakers have access to the precise semantic and syntactic properties of a lexical representation, *Tip-of-the-Tongue (ToT)* phenomena demonstrate that they may nevertheless experience difficulty accessing its phonological form (Brown & McNeill, 1966; Caramazza, 1997). Consequently, speakers may experience uncertainty about both the intended semantics and the phonological form during online production.

To simulate this uncertainty, we generate a "noisy" representation of the production target by introducing random perturbations to the semantic and phonetic representation of the self-repair (w_t^*). We then measure the communicative utility of alternative lexical representations against this noisy version of the target ($\hat{w}_t$). Conceptually, this allows us to model uncertainty about the speaker's intended lexical target rather than assuming that the observed self-repair corresponds exactly to that target. Methodologically, it also enables us to include the repair as an alternative in the model, since the distance between the observed repair and the noisy target representation will generally be non-zero by construction.

Semantic representations are estimated by mapping words to real-valued word embedding vectors. We derive these vectors by fitting `fasttext` embeddings (Bojanowski et al., 2016) to the CANDOR corpus, which yields 100-dimensional semantic vector representations. To generate a noisy semantic representation of the target, Gaussian noise is injected into the target word vector as follows:

$$\hat{w}_t = w_t^* + \boldsymbol{\epsilon}, \tag{8}$$

where $w_t^* \in \mathbb{R}^{100}$ is the vectorized representation of the target and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.1) \in \mathbb{R}^{100}$ is the noise vector. We then operationalize the semantic distance as the cosine distance between the noisy target vector $\hat{w}_t$ and a given word w_i.

We generate a "noisy" phonetic representation of the target by uniformly sampling phonemes from the IPA form of the target word, and for each selected phoneme, uniformly sampling a phonetic feature to modify. For each chosen feature, we then randomly sample an alternative categorical value to add uncertainty to the target phonetic representation (see Appendix H for an algorithmic implementation of this process). Finally, phonetic distance was operationalized as the feature-wise distance between the phonetic representation of a given word w_i and the noisy phonetic representation of the target $\hat{w}_t$. We use the panphon package in Python (Mortensen et al., 2016) to generate categorical phonetic feature representations of words and to compute the feature-based edit distance between the phonetic representation of a given word and the noisy target. See Figure 8 for an example of how these processes generate noisy semantic and phonetic representations of the production target.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

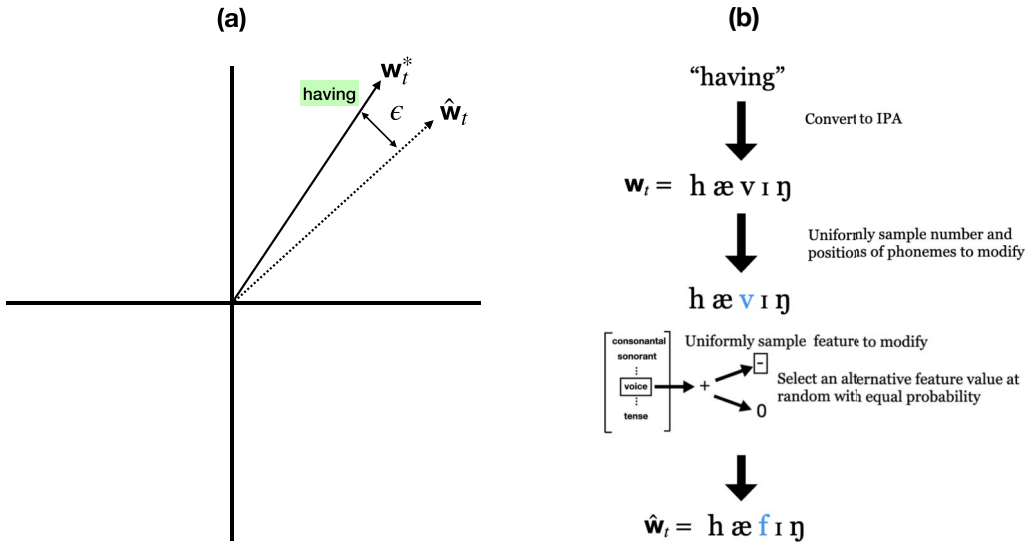


Figure 8. An illustration of the process for computing: (a) the noisy semantic target ($\hat{w}_t$) by injecting noise ($\boldsymbol{\epsilon}$) into the target word's embedding vector ($\hat{w}_t$*) and (b) the noisy phonetic representation of the target using a procedure for injecting noise into the phonetic feature-based target representation. See also Appendix H for a detailed implementation of (b).

The noise process detailed in Figure 8b may yield phonetic representations of the target that do not reflect phonemic, phonotactic, or grammatical category constraints observed in experimentally elicited or naturalistic form-related errors (Baars et al., 1975; Fay & Cutler, 1977; Goldrick et al., 2010). However, we treat this process as analogous to injecting Gaussian noise into the semantic embedding of the target, which may produce vectors that do not correspond to attested lexical items. As with the noisy representation of the target semantics, our goal is simply to introduce random perturbations into the phonetic form to approximate uncertainty about the intended production target.

Regression Methodology. For every utterance, the observed substitution error (e.g., becoming in Ex. 9) is treated as an instance of the positive class whereas all the other words in the vocabulary (N = 14116, including the self-repair, e.g., having in Ex. 9) constitute the negative class. We use the lme4 package to fit the generalized linear model below:

Produced(w_i) ~ Log Unigram Predictability + Log Forward predictability + (Noisy) Semantic Distance + (Noisy) Phonetic Distance

where w_i is a given word in the vocabulary, *Produced*(w_i) = 0 if the word is not the observed substitution error (i.e., if the word is an alternative word from the vocabulary or the self-repair itself), and *Produced*(w_i) = 1 if w_i is the observed error. Put differently, the model learns the weights that best predict the observed lexical mis-selection in context. A positive model coefficient indicates that an increase in the predictor leads to an increase in the log-odds of a word w_i being the observed substitution. That is, it facilitates erroneous production of the word. By comparison, a negative coefficient indicates that an increase in the value of the predictor reduces the log-odds of a word being the observed error, hence inhibiting its production *as an error*. Similar to Study 1, we generate three variants of this model with (i) relative backward predictability (Model 2a), (ii) conditional PMI (Model 2b), and (iii) both relative backward predictability and conditional PMI (Model 2c) to examine the unique and redundant contributions of these measures. See Appendix F.1 for specifications of all the logistic regression models used in these analyses.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

Results

Factors Affecting Substitution Choices. All log-odds coefficients from Models 2a and 2b are presented in Appendix F.2 and all predictors emerged significant. Here we visualize the effects of frequency, predictability, and communicative reward in Figure 9a. In both model variants, unigram predictability was associated with increased log-odds of a word being selected as the observed substitution, with forward predictability showing a similarly consistent positive effect. In contrast, both semantic and phonetic distances exhibited inverse effects: greater distance from the target representation reduced the likelihood of a word being produced as an error. Intriguingly, unlike unigram and forward predictability, relative backward predictability and conditional PMI exhibited negative effects, aligning qualitatively with the distance-based measures. Increased predictability from the future was associated with decreased log-odds of a word being selected as the substitution, both for relative backward predictability and conditional PMI.

Model Comparison. Similar to study 1, we also conduct incremental model comparisons with log-likelihood as a measure of explanatory power. Adding relative backward predictability to the baseline model with unigram predictability, forward predictability, semantic, and phonetic distances led to an improvement in model fit, as evinced by an increase in the log-likelihood ($\Delta\texttt{LogLik} = 13.812$, $\chi^2 = 27.624$, $p < 0.001$). Similarly, adding conditional PMI to the baseline model also improved its goodness of fit ($\Delta\texttt{LogLik} = 20.656$, $\chi^2 = 41.312$, $p < 0.001$). A comparison between these two models revealed that the model with conditional PMI emerged a better model of substitution choices than the model with relative backward predictability ($\Delta\texttt{LogLik} = 6.843$, $\chi^2 = 13.688$, $p < 0.001$). Contrary to the results of Study 1, a comparison between the better model (i.e., model with conditional PMI) and the model with both relative backward predictability and conditional PMI revealed no significant improvements in the substitution model's explanatory power ($\Delta\texttt{LogLik} = 0.845$, $\chi^2 = 1.689$, $p > 0.05$).

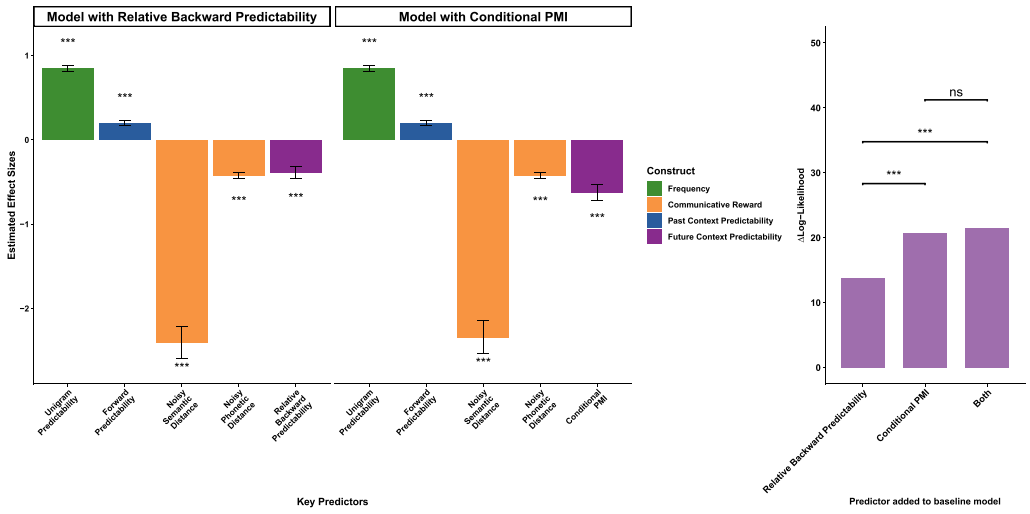


Figure 9. (a) Log-odds coefficients from logistic regression models with relative backward predictability and conditional PMI as operationalizations of future context predictability. Relative backward predictability assumes independence between past and future, whereas conditional PMI assumes conditional dependence. Error bars denote standard error. Positive coefficients indicate that an increase in the predictor facilitates mis-selection of a given word, while negative coefficients inhibit its production as an error. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (*ns*); (b) Δlog-likelihood values obtained from adding future context predictability measures to the baseline model incrementally. Higher values of **ΔLog-Likelihood** indicate a better fit to the data.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

Discussion

The aim of the above analysis was to explicate the distinct contributions of lexical, semantic, and contextual factors in word choice in naturalistic substitution contexts. Across utterances, we observe a strong frequency effect consistent with prior findings: higher-frequency words were more likely to surface as errors than the lower-frequency targets that speakers subsequently produced as self-repairs (Dell, 1990; Kapatsinski, 2010; Kittredge et al., 2008; Shriberg, 1996). Beyond frequency, intrusions were jointly shaped by lexical accessibility and proximity to the target: words that surfaced as errors tended to be frequent and predictable from the past context, yet also lay closer to the target along both semantic and phonetic dimensions. This tension between semantic alignment and form accessibility has been demonstrated in prior research (Ferreira & Griffin, 2003; Koranda et al., 2022) in highly-controlled single-word production paradigms. Our analysis reproduces this pattern at scale, in naturalistic speech, using distributional and phonetic feature-based estimates of semantic and form-based alignment, respectively.

Prior work offers two complementary insights into how contextual predictability shapes production errors. First, priming can boost the accessibility of competing alternatives, increasing the likelihood that they are mis-selected (Bock, 1986; Ferreira & Griffin, 2003; Rapp & Samuel, 2002). Second, less predictable contexts elicit more disfluencies and are associated with higher error rates (Bell et al., 2003; Dammalapati et al., 2019; Fox Tree, 1995). Both of these strands of research concern the *occurrence* of errors. While context is widely assumed to shape lexical selection, to our knowledge this is the first empirical demonstration that both past- and future-context predictability influence. The *identity* of naturalistic substitution errors. Using contextual probabilities derived from neural language models, we show that these effects hold once frequency, semantic alignment, and phonetic alignment are controlled for.

Notably, we find that past and future context appear to exert opposing influences on lexical mis-selection. Like frequency, higher past context predictability increases the likelihood that a word surfaces as an error, suggesting that highly predictable words may be more prone to erroneous production. Future context predictability, however, decreases this likelihood, suggesting that words more compatible with the speaker's future context are less likely to surface as errors.

To illustrate, consider the substitution error in Ex. 11: *So until I see the entire old guard of the Soviet [military/government] completely roll over and disappear...a threat.* Although the past context is compatible with both the error (*military*) and the target (*government*), the future context is substantially more informative about the target (relative backward predictability (*government*) = 1.30, conditional PMI(*government*) = 0.86) than the error (relative backward predictability(*military*) = 0.65, conditional PMI(*military*) = 0.25). More generally, words that share greater informativity with the future context are less likely to surface as errors—an inhibitory effect qualitatively distinct from the facilitative effects of frequency and past predictability on mis-selection. We discuss interpretations of this effect in the General Discussion section.

Both relative backward predictability and conditional PMI emerged as significant predictors of word choice, with larger effect sizes than forward predictability, albeit operating in the opposite direction—inhibiting rather than facilitating error production. Unlike Study 1, the substitution model with conditional PMI outperforms the model with relative backward predictability, and including both predictors yields no improvement in fit, suggesting that relative backward predictability is redundant once conditional PMI is accounted for. A potential explanation is that lexical planning recruits the richer contextual representation captured by conditional PMI, which quantifies how informative the future context is about the present word

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

under the constraints of the observed past. Consequently, relative backward predictability—which ignores the specific past context in which both the future context and the current word are planned—may add little explanatory value once the more contextualized conditional PMI is included.

GENERAL DISCUSSION

Online production is highly, but not strictly, incremental (Ferreira, 1996; Ferreira & Dell, 2000; Ferreira & Swets, 2002; Levelt, 1989; Slevc, 2011; Wagner et al., 2010; Wheeldon et al., 2006): the order in which constituents are planned need not reflect the order in which they are produced (Bock & Levelt, 1994; Ferreira, 2000; Ferreira & Slevc, 2007; Lee et al., 2013; Momma & Ferreira, 2019). Consequently, speaker choices are shaped by predictability from both previously produced past context and planned future context. Prior work has operationalized the latter using backward predictability—the conditional probability of the current word given some future context. However, much of our current understanding of this backward predictability effect comes from corpus studies involving *N*-gram and LSTM models (Bell et al., 2003, 2009; Dammalapati et al., 2021; Harmon & Kapatsinski, 2021; Ranjan et al., 2022), which condition on limited context and can produce coarser probability estimates.

Intriguingly, many of these studies have noted stronger effects of backward predictability on symptoms of production difficulty, such as prolonged pronunciation and disfluency. However, the interpretation of backward predictability both as a probabilistic variable and as a cognitive construct has been complicated by a few factors. First, backward predictability is often highly correlated with forward predictability, which makes it difficult to disentangle their distinct contributions on speaker choices. More importantly, backward predictability captures the predictive effects of future context on the present word, but ignores the context in which both of them occur. Therefore, it can be viewed as a *decontextualized* estimate of how future representations shape the speaker's current choices, since it abstracts away from the particular past context in which those choices are situated.

This work presents an examination of contextual predictability effects, both from the past and future, on two aspects of speaker choice: phonetic encoding and word choice. In addition to clarifying the distinct contributions of past and future context on how speakers choose and encode words, our work addresses the aforementioned limitations with current ways of estimating the predictive effects of future context. We present two alternatives to backward predictability that address issues of collinearity with forward predictability: relative backward predictability, which captures how much more predictable a word is from the future context compared to the past context, and conditional PMI, which quantifies the information gained from the future once the past is taken into account. Unlike backward predictability, conditional PMI measures the unique contribution of the future within that past context, rather than independently of it.

Effects of Predictability on Wordform Encoding

Our first study focused on the effects of predictability on word durations in naturalistic speech. Previous work has established a robust inverse relationship between predictability and wordform variation: predictability from either direction correlates with reduction in phonetic, acoustic, and articulatory detail. Predictability from the future, in particular, has been observed to have a stronger effect on duration than predictability from the past, although previous work has shown that this effect appears to be modulated by lexical class (Bell et al., 2009 cf. Ranjan et al., 2022). In our first study, we find that both proposed alternatives to backward

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

predictability—relative backward predictability and conditional PMI—not only demonstrate this inverse relationship but also emerge as stronger predictors of word duration than forward predictability, in accordance with previous findings. Furthermore, conditional PMI, which captures joint effects of past and future context, explains variance in word durations even alongside relative backward predictability. This finding suggests that such joint effects also shape duration, above and beyond the independent contributions of past and future—which, to our knowledge, has not been demonstrated before.

Previously, Bell et al. (2009) reported asymmetric effects of frequency and contextual predictability on articulatory reduction: whereas frequency and backward predictability emerged as significant predictors of content word duration, forward predictability only affected the duration of high-frequency function words. One possible explanation for the lack of a forward predictability effect on content word durations put forth by Bell and colleagues was that preceding words may be less informative about word choice than following words. Crucially, Bell and colleagues reported no effects of contexts larger than the preceding or following word.

A summary of our key results vis-à-vis Bell et al. (2009)'s findings is presented in Table 2. First, we observe significant predictability effects for context windows that span longer than the preceding or following word, consistent with the improved predictive power of neural language models that condition on broader context than bigram models.

Among the three sources of predictability, we find that future context predictability emerges as the strongest predictor of duration, regardless of operationalization. Furthermore, the present study did not replicate the asymmetric predictability effects of past and future context on function versus content word durations, as in Bell et al. (2009). Predictability from either direction led to reduction in function and content words alike, which suggests that the lexical class distinction may reflect an availability-based effect rather than a separate mechanism of access for function versus content words (cf. Gordon & Dell, 2003).

Yet notably, both function and content word durations were more strongly modulated by predictability from the future context, regardless of the operationalization (i.e., relative backward predictability or conditional PMI). Pluymaekers et al. (2005) interpret stronger predictive effects of the following word on the acoustic realization as evidence of continuous and anticipatory articulatory planning: stronger backward associations enable speakers to compress the production of the current word to allocate resources toward planning and articulating the next one.

A more general interpretation of this effect is that words that are highly predictable from the future context may be produced more quickly, allowing speakers to transition to the production of upcoming words, hence reducing the duration for which these words must be maintained in working memory (Ferreira & Dell, 2000; Slevc, 2011; Wheeldon et al., 2006).

Table 2. A comparison of predictability effects for High-Frequency (HF) function, Mid-Frequency (MF)/Low-Frequency (LF) function, and content words observed in Bell et al. (2009) and the current study.

	Bell et al. (2009)			Current Study		
	HF Function	**MF/LF Function**	**Content**	**HF Function**	**MF/LF Function**	**Content**
Frequency	*ns*	*ns*	*hs*	$p < 0.001$	$p < 0.001$	$p < 0.001$
Predictability from the past	*hs*	*ms*	*ns*		$p < 0.001$	$p < 0.001$
Predictability from the future	*ns*	*hs*	*hs*	$p < 0.001$	$p < 0.001$	$p < 0.001$

Significance as reported by Bell et al. (2009): *ns* = not significant, *ms* = marginally significant, and *hs* = highly significant.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

Table 3. A comparison of the explanatory power of relative backward predictability, conditional PMI, and both predictors combined across the two dependent variables examined in this study.

	Relative Backward Predictability	Conditional PMI	Combined
Word Duration	Stronger predictor	Weaker predictor	Non-redundant; combined model shows best fit
Substitution Choice	Weaker predictor	Stronger predictor	Redundant; no improvement over conditional PMI

Non-redundancy indicates that including both predictors yields a significant improvement in model fit over either predictor alone.

Conversely, when upcoming material is difficult to retrieve, speakers may 'buy time' for planning by prolonging the duration of the current word (Bell et al., 2003; Fox Tree & Clark, 1997).

Although Bell et al. (2009) also acknowledge this anticipatory planning interpretation of the backward predictability effect, they frame predictability effects more generally as arising from the need to coordinate between planning and articulation: when the current word is unpredictable given the past or future context, the production system responds to this processing difficulty by slowing down and prolonging the articulation of the current word. Notably, this view treats predictability from the past and future as mirror effects. Empirically, however, future context predictability consistently emerges as the stronger predictor of word duration, not only in Bell et al. (2009) but also in the present study, which examined multiple operationalizations of this effect. A possible explanation for the stronger influence of future context on durations is that this effect may capture two sources of difficulty: the predictability of a word given some representation of the future sequence, but also the availability of the future representation itself.

Finally, we address the intriguing finding that relative backward predictability—a measure derived from backward predictability which marginalizes over the past context—emerges as a stronger predictor of word duration than conditional PMI, which by definition captures the informativity of the future context given what has already been produced. Yet that these two measures are non-redundant in explaining duration: including both yields a significant improvement in model fit, suggesting they capture distinct aspects of the relationship between future context and phonetic encoding.

We speculate that the marginalization of past context in relative backward predictability yields a decontextualized and aggregated estimate of future context predictability that may provide a better fit to duration, which is a composite phonetic measure that itself integrates multiple bands of planning within the production system. By contrast, conditional PMI likely isolates a more context-specific band of planning that relies on richer representations of past and future context (see Table 3). This may explain why conditional PMI, despite showing an attenuated effect compared to relative backward predictability, still survives as a non-redundant predictor of duration. This interpretation is also consistent with our findings from Study 2, which show that when the dependent variable implicates a more context-sensitive choice i.e., the identity of the substitution error, conditional PMI subsumes relative backward predictability entirely.

The Role of Context in Lexical (Mis-)Selection

The core claim of speaker-oriented accounts of probabilistic reduction is that predictability facilitates lexical access, which, in turn, leads to reduced effort in the downstream encoding of the word. Yet the effects of contextual predictability on the *content* of speaker choices remain underexplored in naturalistic speech, owing to the difficulty of inferring the intended lexical meaning that the speakers

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

aim to convey through their choice of lexical representation. In our second study, we examined the effects of lexical, alignment-driven, and contextual factors on word choice in substitution contexts, where the speaker's self-repair serves as an approximation of their production target.

The core contribution of this analysis, however, is to characterize the role of context in lexical (mis-)selection. Prior work offers two complementary insights into how contextual predictability shapes production errors. First, contextual priming can boost the accessibility of competing alternatives, increasing the likelihood that they are mis-selected (Bock & Warren, 1985; Ferreira & Griffin, 2003; Rapp & Samuel, 2002). Second, less predictable contexts elicit more disfluencies and are associated with higher error rates (Bell et al., 2003; Dammalapati et al., 2019; Fox Tree & Clark, 1997). Both, however, concern the *occurrence* of errors. To our knowledge, this is the first empirical demonstration that predictability from both past and future context, estimated from the same language model, shapes which word, out of a large lexicon of alternatives, surfaces as a naturalistic substitution error. Notably, our modeling of substitution identity shows that once frequency as well as semantic and form-based alignment are controlled for, past and future context predictability exert opposing influences. Like frequency, higher past context predictability increases the likelihood that a word surfaces as an error, suggesting that more predictable words are also more error-prone. Future context predictability, however, shows the opposite effect: words that are more compatible with the speaker's future context appear *less likely* to surface as errors.

In the present study, we adopt a computational-level perspective, focusing on the informational factors that operate during lexical selection. That is, we explicate our findings within a view of lexical selection that is agnostic to assumptions about the nature of lexical representation and stages (or lack thereof) in processing. Instead, we view the lexical (mis-)selections in naturalistic substitutions as resulting from competition between form-, meaning-, and context-based information. While self-repairs may reflect a more optimal prioritization of these informational factors, substitution errors reveal suboptimal weighting of these cues under intrinsic constraints on information processing (Bock, 1982; Ferreira & Pashler, 2002; Futrell, 2023).

A fine-grained categorization of the naturalistic substitutions revealed that roughly 85% of the substitutions in our analysis could be considered lexico-semantic rather than morphosyntactic or phonological. Consequently, the patterns reported here can be viewed as primarily characterizing this class of errors. Across substitutions, word choice reflected a trade-off between speaker effort and communicative reward: words that surfaced as intrusions tended to be frequent and predictable from the past context, while also reflecting a strong semantic and phonological alignment with the target. Beyond this trade-off, we observed a strong inhibitory effect of future context predictability: words more informative about the speaker's upcoming material were less likely to surface *in error*—an effect qualitatively distinct from the facilitatory influence of frequency and past context predictability.

One possible interpretation of this inhibitory effect is that these substitutions arise not from a failure to incorporate future context altogether, but from insufficient integration of future representations into the current selection process. Under this view, while speakers may plan upcoming material in advance (Lee et al., 2013; Momma & Ferreira, 2019; Momma et al., 2016, 2018; Sauppe, 2017), the degree to which pre-planned representations constrain current lexical selection may vary, as actively maintaining and integrating future plans imposes a further burden on verbal working memory (Christiansen & Chater, 2016; Ferreira & Dell, 2000; Slevc, 2011; Wagner et al., 2010). When this integration is suboptimal, future representations may fail to sufficiently constrain lexical selection, allowing a more accessible but contextually inconsistent alternative to slip through as an error.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

An alternative yet complementary interpretation is that predictability from the future may serve as a mismatch signal that flags the inadequacy of the error: when the future context is highly informative about the target but inconsistent with the error, the production system may detect this mismatch and initiate repair, thereby reducing the likelihood that the error surfaces as a substitution. While this interpretation does not directly explain the inhibitory effect of future context on substitution choice, it makes a testable prediction—that greater informativity between the future context and the intended word should facilitate faster error resolution—which we leave to future work. Regardless of the mechanistic interpretation, we find that future context predictability qualitatively differs from past context predictability in that it inhibits rather than facilitates mis-selection—suggesting that past and future context exert opposing pressures on lexical choice, reflecting goal-invariant and goal-directed aspects of lexical planning, respectively.

This balance between goal-invariant or goal-driven processing is captured by the idea of *Good-Enough (GN)* choices in production (Ferreira & Griffin, 2003; Goldberg & Ferreira, 2022; Koranda et al., 2022). Previously, Koranda et al. (2022) demonstrated a trade-off between message alignment and lexical accessibility in word choice: given a speed-accuracy trade-off and multiple possible alternatives to describe a production target, speakers produced the less precise but more frequently experienced alternatives over the more precise yet less accessible choice. Our study builds on this work in two important ways. Whereas Koranda et al. (2022) contrasted the effects of lexical frequency and semantic alignment in a controlled single-word production paradigm, our study extends the scope to word choice in naturalistic utterances, which enables us to examine the role of context in word choice errors. Second, we adopt a comprehender-oriented notion of communicative reward, which encompasses both semantic alignment and proximity to the phonological form of the target. The latter stems from the assumption that comprehenders can infer the intended form through rational inference (Gibson et al., 2013; Ryskin et al., 2018), particularly in the case of errors that are phonologically similar to the target (Upadhye et al., 2025). In other words, errors that are more likely to be correctly inferred by the comprehender may be less costly, and therefore, more communicatively robust compared to those that the interlocutor may find difficult to correct.

Under this expanded notion of good-enough production, substitution errors can be understood as reflecting a suboptimal weighting of informational factors under intrinsic processing constraints. More broadly, this framing aligns with a *resource-rational* view of choice behavior, according to which agents operate under intrinsic cognitive capacity limitations that effectively constrain the information processing that may be used to select the utility-maximizing choice (Anderson, 1991; Gershman et al., 2015; Gigerenzer & Todd, 2008; Griffiths et al., 2015; Lieder & Griffiths, 2020; Simon, 1956). A *rational* response to these constraints is that agents are incentivized to arbitrate between easy-first, heuristic, or automatic, and effortful, goal-driven, or controlled modes of processing, which can lead to choices that reflect a trade-off between utility and information processing costs (Evans & Stanovich, 2013; Futrell, 2023; Kahneman, 1984; Lai & Gershman, 2021; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1984).

Relationship Between Contextual Predictability and Theories of Sentence Planning

Finally, we discuss the link between contextual predictability and theories of planning in sentence production. It is widely assumed that sentence production proceeds through conceptual, functional, positional, and articulatory-motor stages of processing, with 'planning'

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

unfolding at each of these levels of representation (Bock & Levelt, 1994; Ferreira & Slevc, 2007; Hickok, 2012). Under this framework, planning at the conceptual stage involves selecting the relevant semantic, relational, and communicative features of the pre-verbal message; at the functional level, it encompasses both lexical selection and hierarchical structure-building; at the positional level, planning involves morphological encoding and serialization; and at the articulatory-motor level, it involves generating and executing the phonetic plan of the speaker's utterance. These planning processes are assumed to proceed sequentially yet incrementally. As discussed in the preceding section, incrementality in language production emerges as a response to both cognitive constraints and conversational pressures that limit the extent to which speakers can plan their utterance prior to the onset of production (Clark & Wilkes-Gibbs, 1986; Ferreira, 1996; Ferreira & Dell, 2000; Ferreira & Swets, 2002; Kempen & Hoenkamp, 1987; Levelt, 1989; Levinson & Torreira, 2015; Wheeldon et al., 2006 *inter alia*).

However, an outstanding issue in theories of sentence production concerns not just the scope, but also the order in which constituents are planned. A **strictly linear** view of sentence planning maintains that the order of planning roughly mirrors the order in which words appear in the surface form of the utterance (Christiansen & Chater, 2016; De Smedt, 1996; Gleitman et al., 2007; Griffin, 2001; Iwasaki, 2011; Kempen & Hoenkamp, 1987; Levelt, 1989). This strategy exemplifies an *economy of effort* principle in planning: by prioritizing the retrieval and encoding of words that are accessible[12] (Bock & Warren, 1985; Gleitman et al., 2007; Griffin, 2001; McDonald et al., 1993), and minimizing the lag between retrieval and production to the extent permissible by word order constraints, speakers can leverage the processing benefits offered by accessibility while reducing demands on working memory (Christiansen & Chater, 2016; Ferreira & Dell, 2000; Slevc, 2011).

In comparison, a **hierarchical** view of sentence planning underscores the role of structural relations in determining the order in which words are retrieved, with verbs occupying a privileged status in planning due to their central role in specifying clause structure (Bock & Levelt, 1994; Ferreira, 2000; Momma & Ferreira, 2019; Momma et al., 2016; Sauppe, 2017; Schriefers et al., 1998). Under a strong version of this view, words that are linked in a dependency may be retrieved concurrently, regardless of their respective positions in the surface form of the utterance. However, cross-linguistic empirical evidence favors a weaker version of this hypothesis: speakers appear to flexibly arbitrate between highly sequential and hierarchical or look-ahead planning strategies depending on linguistic and meta-linguistic factors such as argument structure, head directionality, word order constraints, and working memory demands (Kidd et al., 2025; Lee et al., 2013; Momma & Ferreira, 2019; Momma et al., 2016, 2018; Nordlinger et al., 2022; Schriefers et al., 1998; Wheeldon et al., 2013).

Although the present study does not account for how speakers arbitrate between strictly linear and hierarchical planning or generate predictions about the order in which words are planned, these two strategies provide a rationale for the observed effects of past and future context predictability on both word duration and word choice. Strictly linear planning can be viewed as analogous to incremental prediction, wherein the production of previous words facilitates access of contextually predictable continuations, reflecting an accessibility-driven strategy. Hierarchical planning, on the other hand, provides a motivation for the early

[12] As operationalized by various correlates of accessibility such as givenness, imageability, animacy, and subjecthood.

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

availability of future material. Since both the scope and content of a speaker's future plans appear to be sensitive to cognitive load, this planning strategy can be viewed as reflecting more effortful or goal-oriented planning. From a resource-rational perspective, this interleaving of planning strategies is consistent with our findings that past and future context predictability appear to reflect goal-invariant and goal-directed pressures on lexical choice, respectively.

Moreover, this flexibility in planning also provides further motivation for our proposed measure of future context predictability based on conditional PMI(w_t; $C_{>t}|C_{<t}$). Although backward predictability quantifies the predictive effect of future context on the present word, this effect is not situated in the context of what the speaker has already produced. Furthermore, by ignoring any direct information-processing dependencies between words in the past and future contexts, the measure assumes that the future context depends only on the current word. An implication of this assumption that is of relevance to theories of sentence planning is that it assumes that words in the future are planned strictly linearly. By preserving dependencies between past and future contexts and providing a contextualized measure of informativity between the speaker's future plans and the current word, conditional PMI can serve as a more interpretable probabilistic variable for modeling the effects of non-incremental planning.

Moreoever, conditional PMI is a *symmetric* quantity that measures how informative the current word and the future context are about each other in the context of past sequence. In earlier discussion, we focused on a backward-looking interpretation, where conditional PMI captures the constraining effect of an already planned future sequence on present choice. However, this quantity is also compatible with a forward-looking or prospective interpretation of planning such as *forward simulation*, which aligns with how planning is widely conceptualized in action planning, optimal control, and reinforcement learning. Forward simulation, broadly construed, involves generating possible future trajectories of actions according to the agent's policy, conditioned on both the present state (which reflects the outcomes of past actions) and an action sampled at the current time step (Sutton & Barto, 1998). Action selection at the current time-step, therefore, is guided by the *future value* of a candidate action i.e., the expected return aggregated over the simulated future trajectories (Todorov, 2009). In other words, agents choose the current action with the objective of facilitating a future trajectory that maximizes expected utility or communicative value. In the context of sentence planning, this prospective interpretation of conditional PMI can be viewed as speakers choosing words that are informative and facilitate production of any upcoming material. Since the present study assumes the entire future sequence as a single deterministic chunk, our findings are compatible with both interpretations. We leave it to future work to examine how uncertainty in the representation of the future sequence may affect how speakers choose and encode words.

CONCLUSION

This work examines the role of past and future context on wordform encoding and word choice in naturalistic language production. The effect of future context or backward predictability, in particular, has long remained understudied compared to the effects of past context or forward predictability. Beyond the methodological and statistical confounds linked to language modeling and multi-collinearity, the link between backward predictability and mechanisms of sentence planning is obscured by the assumption that future context is statistically independent of the past context. In this work, we introduce a principled alternative to backward predictability based on the conditional pointwise mutual information between the current word and the future sequence conditioned on the previously produced context. Our

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

empirical contributions are two-fold. First, we revisit findings in probabilistic reduction, both as a validation of our proposed alternative and to re-evaluate earlier claims in light of improved measures of contextual predictability. Broadly, results from our first study reveal that predictability from the future exerts a stronger influence on word duration regardless of lexical class, which suggests uncertainty about the future context may further incentivize speakers to slow production. In our second study, we present a model of substitution errors in context, which explicates the unique contributions of availability-based preferences, communicative utility, and context on word choice. Beyond replicating trade-offs between lexical availability and communicative utility previously observed in single-word production, our findings reveal that past and future context exert distinct, and indeed opposing, influences on lexical choice. Overall, our study helps bridge psycholinguistic theories of sentence production with probabilistic and resource-rational approaches to modeling the mechanisms that underlie language production.

ACKNOWLEDGMENTS

We thank the editor and two anonymous reviewers for their careful and constructive feedback, which substantially improved this manuscript. We also thank members of the Language Processing Group and Quantlang Lab at UC Irvine and the Language Production Lab at UCSD for their feedback. This work benefited from presentation at the InfoCog workshop at NeurIPS 2022 and the 47th Annual Meeting of the Cognitive Science Society. We thank Frankie Boren for her assistance in data annotation, Dr. Shota Momma for helpful discussion, and Dr. Cassandra Jacobs for valuable correspondence regarding the function-content coding in our dataset.

FUNDING INFORMATION

This work was supported by a National Science Foundation grant (#1947307) awarded to R.F.

AUTHOR CONTRIBUTIONS

S.U.: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review & editing. R.F.: Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – review & editing.

DATA AND CODE AVAILABILITY STATEMENT

Data and analysis scripts used in this study are available at https://github.com/shiupadhye/past-future-predictability-production.

REFERENCES

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(Pt 1), 31–56. https://doi.org/10.1177/00238309040470010201, PubMed: 15298329

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, *14*(4), 382–391. https://doi.org/10.1016/S0022-5371(75)80017-X

Balota, D. A., Boland, J. E., & Shields, L. W. (1989). Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language*, *28*(1), 14–36. https://doi.org/10.1016/0749-596X(89)90026-0

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

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

Bavarian, M., Jun, H., Tezak, N., Schulman, J., McLeavey, C., Tworek, J., & Chen, M. (2022). Efficient training of language models to fill in the middle. *arXiv*. https://doi.org/10.48550/arXiv.2207.14255

Beckman, M. E., & Hirschberg, J. (1994). *The Tobi annotation conventions*. Ohio State University.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, *60*(1), 92–111. https://doi.org/10.1016/j.jml.2008.06.003

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *Journal of the Acoustical Society of America*, *113*(2), 1001–1024. https://doi.org/10.1121/1.1534836, PubMed: 12597194

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*(1), 1–47. https://doi.org/10.1037/0033-295X.89.1.1

Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 575–586. https://doi.org/10.1037/0278-7393.12.4.575

Bock, J. K., & Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 467–484. https://doi.org/10.1016/S0022-5371(80)90321-7

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*(1), 47–67. https://doi.org/10.1016/0010-0277(85)90023-X, PubMed: 4075761

Bock, K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). Academic Press.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv*. https://doi.org/10.48550/arXiv.1607.04606

Breaugh, J. A. (2006). Rethinking the control of nuisance variables in theory testing. *Journal of Business and Psychology*, *20*(3), 429–443. https://doi.org/10.1007/s10869-005-9009-y

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 325–337. https://doi.org/10.1016/S0022-5371(66)80040-3

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*(3), 261–290. https://doi.org/10.1017/S0954394502143018

Bybee, J. L., & Hopper, P. J. (2001). *Frequency and the emergence of linguistic structure* (Vol. 45). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.45

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., & Beaver, D. I. (2010). The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, *44*, 387–419. https://doi.org/10.1007/s10579-010-9120-1

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177–208. https://doi.org/10.1080/026432997381664

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X, PubMed: 25869618

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7, PubMed: 3709088

Clark, T. H., Poliak, M., Regev, T., Haskins, A. J., Gibson, E., & Robertson, C. (2025). The relationship between surprisal, prosody, and backchannels in conversation reflects intelligibility-oriented pressures. *Cognitive Science*, *49*(10), e70134. https://doi.org/10.1111/cogs.70134, PubMed: 41144721

Clopper, C. G., &Turnbull, R. (2018). Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, & M. Zellers (Eds.), *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation* (pp. 25–72). De Gruyter Mouton. https://doi.org/10.1515/9783110524178-002

Cutting, J. C., & Ferreira, V. S. (1999). Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 318–344. https://doi.org/10.1037//0278-7393.25.2.318, PubMed: 10093205

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. https://doi.org/10.18653/v1/N19-3015

Dammalapati, S., Rajkumar, R., & Agarwal, S. (2021). Effects of duration, locality, and surprisal in speech disfluency prediction in English spontaneous speech. In *Proceedings of the Society for Computation in Linguistics 2021* (pp. 91–101). Online. Association for Computational Linguistics.

De Smedt, K. (1996). Computational models of incremental grammatical encoding. In T. Dijkstra & K. de Smedt (Eds.), *Computational psycholinguistics: AI and connectionist models of human language processing* (pp. 279–307). Taylor & Francis.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*(4), 313–349. https://doi.org/10.1080/01690969008407066

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*(1–3), 287–314. https://doi.org/10.1016/0010-0277(92)90046-K, PubMed: 1582160

Dell, G. S., Oppenheim, G. M., & Kittredge, A. K. (2008). Saying the right word at the right time: Syntagmatic and paradigmatic interference in sentence production. *Language and Cognitive Processes*, *23*(4), 583–608. https://doi.org/10.1080/01690960801920735, PubMed: 20622975

Demberg, V., Sayeed, A., Gorinski, P., & Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In J. Tsujii, J. Henderson, & M. Paşca (Eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 356–367). Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Donahue, C., Lee, M., & Liang, P. (2020). Enabling language models to fill in the blanks. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2492–2501). Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.225

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685, PubMed: 26172965

Fano, R. M., & Hawkins, D. (1961). Transmission of information: A statistical theory of communications. *American Journal of Physics*, *29*(11), 793–794. https://doi.org/10.1119/1.1937609

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

Fehlauer, F., Mahowald, K., & Pimentel, T. (2025). Convergence and divergence of language models under different random seeds. *arXiv*. https://doi.org/10.48550/arXiv.2509.26643

Ferreira, F. (2000). Syntax in language production: An approach using tree-adjoining grammars. In L. R. Wheeldon (Ed.), *Aspects of language production* (pp. 291–330). Psychology Press.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*(1), 57–84. https://doi.org/10.1006/jmla.2001.2797

Ferreira, V., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1187–1199. PubMed: 12450341

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language*, *35*(5), 724–755. https://doi.org/10.1006/jmla.1996.0038

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*(4), 296–340. https://doi.org/10.1006/cogp.1999.0730, PubMed: 10888342

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, PubMed: 12564760

Ferreira, V. S., Morgan, A. M., & Slevc, L. R. (2018). Grammatical encoding. *The Oxford Handbook of Psycholinguistics* (pp. 432–457).

Ferreira, V. S., & Slevc, L. R. (2007). Grammatical encoding. In G. M. Gaskell (Ed.), *The oxford handbook of Psycholinguistics* (pp. 453–469). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568971.013.0027

Fox, T. J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*(6), 709–738. https://doi.org/10.1006/jmla.1995.1032

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*(2), 151–167. https://doi.org/10.1016/S0010-0277(96)00781-0, PubMed: 9141905

Freedman, D. A. (2009). *Statistical Models: Theory and Practice* (pp. 431–442). Cambridge University Press. https://doi.org/10.1017/CBO9780511815867

Fromkin, V. A. (2000). *Fromkin's speech error database*. Online database. Max Planck Institute for Psycholinguistics, Nijmegen. https://www.mpi.nl/resources/data/fromkins-speech-error-database/

Futrell, R. (2023). Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, *120*(39), e2220593120. https://doi.org/10.1073/pnas.2220593120, PubMed: 37725652

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814. https://doi.org/10.1111/cogs.12814, PubMed: 32100918

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *bioRxiv*.

Gershman, S. J., & Bhui, R. (2020). Rationally inattentive intertemporal choice. *bioRxiv*. https://doi.org/10.1038/s41467-020-16852-y

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. https://doi.org/10.1126/science.aac6076, PubMed: 26185246

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110, PubMed: 23637344

Gigerenzer, G., & Todd, P. M. (2008). Rationality the fast and frugal way: Introduction. *Handbook of Experimental Economics Results*, *1*, 976–986. https://doi.org/10.1016/S1574-0722(07)00105-9

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(4), 544–569. https://doi.org/10.1016/j.jml.2007.01.007, PubMed: 18978929

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 517–520). IEEE Computer Society. https://doi.org/10.1109/ICASSP.1992.225858

Goldberg, A. E., & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, *26*(4), 300–311. https://doi.org/10.1016/j.tics.2022.01.005, PubMed: 35241380

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

Goldrick, M., Baker, H. R., Murphy, A., & Baese-Berk, M. (2011). Interaction and representational integration: Evidence from speech errors. *Cognition*, *121*(1), 58–72. https://doi.org/10.1016/j.cognition.2011.05.006, PubMed: 21669409

Goldrick, M., Folk, J. R., & Rapp, B. (2010). Mrs. malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language*, *62*(2), 113–134. https://doi.org/10.1016/j.jml.2009.11.008, PubMed: 20161591

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: An account of deficits in light and heavy verb production. *Cognitive Science*, *27*(1), 1–40. https://doi.org/10.1207/s15516709cog2701_1

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, *82*(1), B1–B14. https://doi.org/10.1016/S0010-0277(01)00138-X, PubMed: 11672707

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. https://doi.org/10.1111/tops.12142, PubMed: 25898807

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Harmon, Z., & Kapatsinski, V. (2021). A theory of repetition and retrieval in language production. *Psychological Review*, *128*(6), 1112–1144. https://doi.org/10.1037/rev0000305, PubMed: 34242049

Hartsuiker, R. J., & Moors, A. (2017). On the automaticity of language processing. In H.-J. Schmid (Ed.), *Entrenchment and the*

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

psychology of language learning: How we reorganize and adapt linguistic knowledge (pp. 201–225). De Gruyter Mouton. https://doi.org/10.1037/15969-010

Hashimoto, D. (2021). Probabilistic reduction and mental accumulation in japanese: Frequency, contextual predictability, and average predictability. *Journal of Phonetics*, *87*, 101061. https://doi.org/10.1016/j.wocn.2021.101061

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*(2), 135–145. https://doi.org/10.1038/nrn3158, PubMed: 22218206

Iwasaki, N. (2011). Incremental sentence production: Observations from elicited speech errors in japanese. In H. Yamashita, Y. Hirose, & J. Packard (Eds.), *Processing and producing head-final structures* (pp. 131–151). Springer, Dordrecht. https://doi.org/10.1007/978-90-481-9213-7_7

Jacobs, C. L., Buxó-Lugo, A., Taylor, A. K., & Leopold-Hooke, M. (2025). Less is more: Probabilistic reduction is best explained by small-scale predictability measures. *arXiv*. https://doi.org/10.48550/arXiv.2512.23659

Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech*. [Doctoral dissertation, Stanford University]. ProQuest Dissertations and Theses Global.

Jaeger, T. F., & Buz, E. (2017). Signal reduction and linguistic encoding. In E. M. Fernández & H. S. Cairns (Eds.), *The handbook of psycholinguistics* (pp. 38–81). John Wiley & Sons. https://doi.org/10.1002/9781118829516.ch3

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). "Probabilistic relations between words: Evidence from reduction in lexical production". In J. L. Bybee & P. J. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 229–254). https://doi.org/10.1075/tsl.45.13jur

Kahneman, D., & Treiesman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D. R. Davis (Eds.), *Varieties of attention* (pp. 29–61). Academic Press.

Kapatsinski, V. (2010). Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and Speech*, *53*(Pt 1), 71–105. https://doi.org/10.1177/0023830909351220, PubMed: 20415003

Kapatsinski, V. (2026). Transformers perform adaptive partial pooling. *arXiv*. https://doi.org/10.48550/arXiv.2602.03980

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, *11*(2), 201–258. https://doi.org/10.1207/s15516709cog1102_5

Kidd, E., Garrido Rodrguez, G., Wilmoth, S., Garrido Guillén, J. E., & Nordlinger, R. (2025). How does speaking a free word order language influence sentence planning and production? Evidence from pitjantjatjara (pama-nyungan, australia). *Cognitive Science*, *49*(7), e70087. https://doi.org/10.1111/cogs.70087, PubMed: 40684369

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology*, *25*(4), 463–492. https://doi.org/10.1080/02643290701674851, PubMed: 18704797

Koranda, M. J., Zettersten, M., & MacDonald, M. C. (2022). Good-enough production: Selecting easier words instead of more accurate ones. *Psychological Science*, *33*(9), 1440–1451. https://doi.org/10.1177/09567976221089603, PubMed: 35942911

Kurumada, C. (2011). Syntactic context affects probability estimation: Evidence from Japanese relative clauses. In *85th LSA*.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In K. D. Federmeier (Ed.), *The Psychology of learning and motivation* (Vol. 74, pp. 195–232). Elsevier Academic Press. https://doi.org/10.1016/bs.plm.2021.02.004

Lashley, K. S. (1951). *The problem of serial order in behavior* (Vol. 21). Bobbs-Merrill Oxford.

Lee, E.-K., Brown-Schmidt, S., & Watson, D. G. (2013). Ways of looking ahead: Hierarchical planning in language production. *Cognition*, *129*(3), 544–562. https://doi.org/10.1016/j.cognition.2013.08.007, PubMed: 24045002

Levelt, W. J. (1981). The speaker's linearization problem. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *295*, 305–315. https://doi.org/10.1098/rstb.1981.0142

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*, 731. https://doi.org/10.3389/fpsyg.2015.00731, PubMed: 26124727

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006, PubMed: 17662975

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. https://doi.org/10.1017/S0140525X1900061X, PubMed: 30714890

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. https://doi.org/10.1016/j.cognition.2012.09.010, PubMed: 23116925

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313–330.

McDonald, J. L., Bock, J. K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, *25*(2), 188–230. https://doi.org/10.1006/cogp.1993.1005, PubMed: 8482072

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the Uniform Information Density hypothesis. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 963–980). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.74

Momma, S., & Ferreira, V. S. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, *114*, 101228. https://doi.org/10.1016/j.cogpsych.2019.101228, PubMed: 31421521

Momma, S., Slevc, L. R., & Phillips, C. (2016). The timing of verb selection in Japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 813–824. https://doi.org/10.1037/xlm0000195, PubMed: 26569434

Momma, S., Slevc, L. R., & Phillips, C. (2018). Unaccusativity in sentence production. *Linguistic Inquiry*, *49*(1), 181–194. https://doi.org/10.1162/LING_a_00271

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. S. (2016). Panphon: A resource for mapping IPA segments to articulatory feature vectors. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical*

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

Papers (pp. 3475–3484). The COLING 2016 Organizing Committee.
Nordlinger, R., Rodriguez, G. G., & Kidd, E. (2022). Sentence planning and production in murrinhpatha, an australian 'Free Word Order' language. *Language*, *98*(2), 187–220. https://doi.org/10.1353/lan.2022.0008
Opedal, A., Chodroff, E., Cotterell, R., & Wilcox, E. (2024). On the role of context in reading time prediction. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 3042–3058). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.179
Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven, & N. Warner (Eds)., *Laboratory Phonology* (Vol. 7, pp. 101–140). De Gruyter Mouton. https://doi.org/10.1515/9783110197105.1.101
Pimentel, T., Meister, C., Salesky, E., Teufel, S., Blasi, D., & Cotterell, R. (2021). A surprisal–duration trade-off across and within the world's languages. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 949–962). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.73
Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2–4), 146–159. https://doi.org/10.1159/000090095, PubMed: 16391500
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., Child, R., Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & Wu, J. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*, 9.
Ranjan, S., Rajkumar, R., & Agarwal, S. (2022). Linguistic complexity and planning effects on word duration in Hindi read aloud speech. In A. Ettinger, T. Hunter, & B.Prickett Prickett (Eds.), *Proceedings of the Society for Computation in Linguistics 2022* (pp. 119–132). Association for Computational Linguistics.
Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, *107*(3), 460–499. https://doi.org/10.1037/0033-295X.107.3.460, PubMed: 10941277
Rapp, D. N., & Samuel, A. G. (2002). A reason to rhyme: Phonological and semantic influences on lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 564–571. https://doi.org/10.1037//0278-7393.28.3.564, PubMed: 12018508
Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A., & Marin, S. (2023). The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, *9*(13), eadf3197. https://doi.org/10.1126/sciadv.adf3197, PubMed: 37000886
Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, *181*, 141–150. https://doi.org/10.1016/j.cognition.2018.08.018, PubMed: 30195136
Sauppe, S. (2017). Word order and voice influence the timing of verb planning in german sentence production. *Frontiers in Psychology*, *8*, 1648. https://doi.org/10.3389/fpsyg.2017.01648, PubMed: 29018379
Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1
Schriefers, H., Teruel, E., & Meinshausen, R.-M. (1998). Producing simple sentences: Results from picture–word interference experiments. *Journal of Memory and Language*, *39*(4), 609–632. https://doi.org/10.1006/jmla.1998.2578
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, PubMed: 25019178
Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121, PubMed: 38422017
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(4), 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x
Shiffrin, R. M., & Schneider, W. (1984). Automatic and controlled processing revisited. *Psychological Review*, *91*(2), 269–276. https://doi.org/10.1037/0033-295X.91.2.269, PubMed: 6571425
Shriberg, E. (1996). Disfluencies in switchboard. In *Proceedings of International Conference on Spoken Language Processing* (Vol. 96, pp. 11–14). IEEE.
Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769, PubMed: 13310708
Slevc, L. R. (2011). Saying what's on your mind: Working memory effects on sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1503–1514. https://doi.org/10.1037/a0024350, PubMed: 21767058
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.
Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, *1*(2), 147–165. https://doi.org/10.1515/LANGCOG.2009.008
Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, *106*(28), 11478–11483. https://doi.org/10.1073/pnas.0710743106, PubMed: 19574462
Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in american english words. *Journal of Phonetics*, *35*(4), 445–472. https://doi.org/10.1016/j.wocn.2006.12.001
Umair, M., Mertens, J. B., Warnke, L., & de Ruiter, J. P. (2024). Can language models trained on written monologue learn to predict spoken dialogue? *Cognitive Science*, *48*(11), e70013. https://doi.org/10.1111/cogs.70013, PubMed: 39587969
Upadhye, S., Li, J., & Futrell, R. (2025). Spacer: A parallel dataset of speech production and comprehension of error repairs. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, J. Li, & B.-D. Oh (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 144–154). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.cmcl-1.19
Wagner, V., Jescheniak, J. D., & Schriefers, H. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 423–440. https://doi.org/10.1037/a0018619, PubMed: 20192540
Warnke, L., & de Ruiter, J. P. (2023). Top-down effect of dialogue coherence on perceived speaker identity. *Scientific Reports*, *13*(1),

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

3458. https://doi.org/10.1038/s41598-023-30435-z, PubMed: 36859459

Wheeldon, L., Ohlson, N., Ashby, A., & Gator, S. (2013). Lexical availability and grammatical encoding scope during spoken sentence production. *Quarterly Journal of Experimental Psychology*, *66*(8), 1653–1673. https://doi.org/10.1080/17470218.2012.754913, PubMed: 23286440

Wheeldon, L. R., Meyer, A. S., & Smith, M. (2006). Language production, incremental. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science*. John Wiley & Son. https://doi.org/10.1002/0470018860.s00252

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*(3), 1707–1717. https://doi.org/10.1121/1.402450, PubMed: 1564206

Wilcox, E. G., Gauthier, J., & Hu, J. (2020). On the predictive power of neural language models for human real-timecomprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 42).

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612

Wolf, L., Pimentel, T., Fedorenko, E., Cotterell, R., Warstadt, A., Wilcox, E. G., & Regev, T. I. (2023). Quantifying the redundancy between prosody and text. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9765–9784). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.606

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48. https://doi.org/10.1016/j.jml.2013.12.003

Xu, W., Chon, J., Liu, T., & Futrell, R. (2023). The linearity of the effect of surprisal on reading times across languages. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15711–15721). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1052

Zhan, M., & Levy, R. P. (2018). Comparing theories of speaker choice using a model of classifier production in mandarin chinese. 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, Volume 1 (Long Papers)* (pp. 1997–2005). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-11811

Downloaded from direct.mit.edu/opmi/article-pdf/doi/10.1162/OPMI.a.397/2630536/opmi.a.397.pdf by guest on 11 November 2026

A Data and Code Availability

Data and analysis scripts used in this study are available at https://github.com/shiupadhye/past-future-predictability-production.

B Derivation of Conditional PMI from Value-to-go

As noted in Section 2.1, conditional PMI is a symmetric quantity that does not commit to a causal planning mechanism. Rather, it captures how informative the current word (w_t) and future sequence ($C_{<t} = \{w_{t+1}, \cdots w_N\}$) are about each other in the context of what the speaker has already produced ($C_{>t} = \{w_1, \cdots w_{t-1}\}$). Therefore, it is consistent with both a *retrospective* interpretation, where material retrieved in advance of production constrains word choice, and a *prospective* interpretation in which the current word is chosen such that it aids in the production of desired future words.

Applied to lexical planning, the control-theoretic concept of *value-to-go* can be viewed as the expected value or reward of choosing a word w_t, which includes both the value of choosing the current word and expected value of a future sequence of words ($c = w_{t+1} \cdots w_N$):

$$\mathbb{E}_{c \sim p(.|w_t, C_{<t})}[v_M(c \mid w_t, C_{<t})] = \sum_c v_M(c \mid w_t, C_{<t}) \;\; p(c \mid w_t, C_{<t}) \tag{9}$$

Eq. 9 denotes the expected future value of choosing a word at time t by computing the expectation or aggregating over all possible future sequences. Here, v_M denotes the *communicative value* of a future (c) under the speaker's goal or message M and conditioned on the choice of the current word and the observed past context.

If we want to use this notion of value-to-go as a predictor of speakers' word choices, we are faced with a problem: enumerating all possible future sequences to compute the sum in Eq. 9 is highly intractable. Therefore, we make three simplifying assumptions about the possible future context. First, we treat the entire future context as a **one** multi-word chunk or action taken at time-step $t+1$. Furthermore, for simplicity, we consider only the future context actually produced by the speaker $C_{>t}$ instead of computing an expectation over all possible alternative futures. Finally, we hold the communicative value (v_M) constant for all choices of w_t. In other words, word choice at time t does not change the communicative value of the observed future, which we already treat as fixed with respect to M.

These assumptions enable us to simplify Eq. 9 to the probability of the future conditioned on a selected word w_t at time t and observed past $C_{<t}$:

$$p(C_{>t} \mid w_t, C_{<t}) \tag{10}$$

Below, we show that conditional PMI$(C_{>t}; w_t \mid C_{<t})$ can be derived from Eq. 10 as follows:

Applying Bayes rule, we can rewrite $p(C_{>t} \mid w_t, C_{<t})$ in terms of the bidirectional probability $p(w_t \mid C_{>t}, C_{<t})$, forward word probability $p(w_t \mid C_{<t}$ and the conditional probability of the $p(C_{>t} \mid C_{<t})$:

$$p(C_{>t} \mid w_t, C_{>t}) = \frac{p(w_t \mid C_{>t},\ C_{>t})\ p(C_{>t} \mid C_{<t})}{P(w_t \mid C_{<t})} \tag{11}$$

Applying log-transformation to Eq. 11

$$\log p(C_{>t} \mid w_t, C_{>t}) = \log p(w_t \mid C_{>t},\ C_{<t})\ - \log p(w_t \mid C_{<t})\ + \log p(C_{>t} \mid C_{<t}) \tag{12}$$

Note that $p(c \mid C_{<t})$ is a constant with respect to w_t. Hence, we can further simplify Eq. 12 as follows:

$$\begin{aligned} &\propto \log p(w_t \mid C_{>t},\ C_{<t})\ - \log p(w_t \mid C_{<t}) \\ &= \text{Conditional PMI}(C_{>t}; w_t \mid C_{<t}) \end{aligned} \tag{13}$$

C Language Model Training and Evaluation

A GPT-2 `small` transformer language model was pre-trained using the data augmentation process detailed in Figure 5. Instead of using an off-the-shelf Byte-Pair Encoding (BPE) tokenizer (Radford et al., 2019), which relies on subword tokenization based on token frequency, we opt to train a whitespace-based tokenizer on the CANDOR corpus; we choose this tokenization scheme for simplicity of estimating token or word probability and to keep the vocabulary size more tractable (N = 14116). Beyond this, the tokenizer vocabulary also included `<eos>`, `<PRE>`, `<SUF>`, `<MID>`, and `<unk>` tokens. Below, we provide an algorithmic implementation of the process for augmenting the model training and evaluation datasets.

Algorithm 1: Procedure for augmenting training and evaluation datasets for infill estimation

Input: Corpus U

Output: Augmented corpus U'

for $u \in U$ **do** `// Iterate over every utterance in the corpus`

- $N \leftarrow length(u)$
- $k \sim \mathcal{U}(1, N)$ `// Uniformly sample the position of a word`
- $\rho \sim Bernoulli(0.5)$
- **if** *prefix precedes suffix in* u' **then**
  - $\rho = 0$
  - $u' \leftarrow w_1 \cdots w_{k-1} w_{k+1} \cdots w_N w_k$ `// Prefix precedes suffix in` u'
- **else**
  - $u' \leftarrow w_{k+1} \cdots w_N w_1 \cdots w_{k-1} w_k$ `// Suffix precedes prefix in` u'

Model hyperparameters were determined via grid-search and performance was evaluated based on language modeling perplexity. The finalized hyperparameters are summarized in Table 4.

Hyperparameter	**Value**
Context Window	1024
Initial Learning Rate	5×10^{-5}
Training Batch Size	4
Validation Batch Size	4
Epochs	10 (with early-stopping)
L2-regularization	0.01

Table 4: Hyperparameters for training GPT-2 SMALL (124M parameter) language model.

Whereas the model was trained on the CANDOR corpus, the duration and substitution utterances were extracted from Switchboard. It bears mentioning that these spontaneous speech corpora were compiled over three decades apart, and the CANDOR corpus comprises conversations over video chat while conversations in Switchboard were over the telephone. Consequently, we anticipate differences across corpora, which may be due to diachronic trends in language use and modality-specific factors such as audio versus visual feedback during conversation. Therefore, to ensure that the model trained on CANDOR generalized to Switchboard, the models were evaluated on a held-out subset of the Switchboard corpus. The infill-trained GPT-2 achieved a perplexity of 16.195 on Switchboard while forward and backward-trained GPT-2 models achieved perplexities of 39.2778 and 38.3563, respectively. The lower perplexity for the infill-trained model suggests that bidirectional contextual information—even when preceding and following contexts are presented outside of canonical order—leads to improved prediction of the word in context.

Forward and backward predictability estimates from both models were found to be highly correlated, suggesting that infill-trained model could reliably estimate autoregressive probabilities in the forward and backward directions. Correlations between these estimates are presented in Figure 10

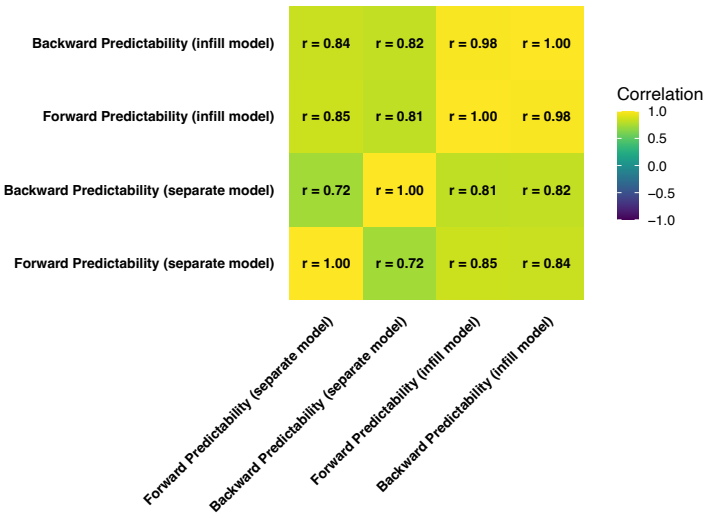


Figure 10: Pairwise Pearson correlations between forward and backward predictability estimates from separate unidirectional and infill-trained models.

D Study 1: Statistical Models and Results

D.1 Linear regressions for modeling word durations

Table 5 details the specifications of the mixed effects regression models that were fit to word durations from the Switchboard corpus in Section 3.

Model	**Model Equation**
Model 1 (Baseline)	Duration ~ Unigram Probability + Forward Predictability + Word Length + Speech Rate + Speaker Age + Speaker Sex + (1 | Speaker)
Model 1a	Duration ~ Unigram Probability + Forward Predictability + **Relative Backward Predictability** + Word Length + Speech Rate + Speaker Age + Speaker Sex + (1 | Speaker)
Model 1b	Duration ~ Unigram Probability + Forward Predictability + **Conditional PMI** + Predictability + Word Length + Speech Rate + Speaker Age + Speaker Sex + (1 | Speaker)
Model 1c (both)	Duration ~ Unigram Probability + Forward Predictability + Relative **Backward Predictability** + **Conditional PMI** + Predictability + Word Length + Speech Rate + Speaker Age + Speaker Sex + (1 | Speaker)
Model 1a (predictability * lexical class)	Duration ~ **Unigram Probability * Lexical Class** + **Forward Predictability * Lexical Class** + **Relative Backward Predictability * Lexical Class** + Word Length + Speech Rate + Speaker Age + Speaker Sex + (1 | Speaker)
Model 1b (predictability * lexical class)	Duration ~ **Unigram Probability * Lexical Class** + **Forward Predictability * Lexical Class** + **Conditional PMI * Lexical Class** + Word Length + Speech Rate + Speaker Age + Speaker Sex + (1 | Speaker)

Table 5: Linear Mixed-Effects Regression models fit using `lme4`. Incrementally added predictors are denoted in **bold**.

D.2 Regression coefficients for analysis of function and content word durations

Regression coefficients from the lexical class analysis in Section 3 are summarized in Table 6.

	Model 1a (with relative backward predictability)	**Model 1b** (with conditional PMI)
(Intercept)	150.646(2.91)***	149.528(2.88)***
Unigram Predictability	-17.456(0.26)***	-17.721(0.26)***
Predictability from the past	2.80(0.25)***	3.497(0.25)***
Predictability from the future	-9.492(0.58)***	-4.376(0.85)***
Lexical Class:Content	1.94(1.56)	-18.136(1.57)***
Word Length (in syllables)	82.127(0.37)***	81.593(0.37)***
Speech Rate	-27.618(0.14)***	-27.943(0.14)***
Speaker Age	0.212(0.06)**	0.218(0.06)***
Speaker Sex:M	-8.911(1.37)***	-8.198(1.35)***
Unigram Predictability x Lexical Class:Content	-3.381(0.28)**	-3.886(0.28)**
Predictability from the past x Lexical Class:Content	-6.793(0.28)***	-6.384(0.28)***
Predictability from the future x Lexical Class:Content	-27.282(2.82)***	-29.67(1.21)***

Table 6: Regression coefficients from models with relative backward predictability and conditional PMI as operationalizations of future context predictability. Probabilistic predictors are bolded. Parentheses denote standard error. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (ns).

E Study 1: Additional Analyses

E.1 Extended Model Comparisons with Backward Predictability and Unconditional PMI

In our main analysis, we focused on two operationalizations of future context predictability: relative backward predictability and conditional PMI. Our rationale for using relative backward predictability is that it is effectively decorrelated from forward predictability and maintains the asymmetry of backward predictability while still yielding a model with equivalent explanatory power. As noted in Section 2.1, unconditional PMI (Eq. 3) also assumes independent effects of past and future context similar to relative backward predictability. Where these two measures differ, however, is that unconditional PMI is a symmetric measure that quantifies the strength of association between the current word and the future. Below, we replicate the analysis conducted in Section 3 with standard backward predictability and unconditional PMI to sketch a complete comparison between different operationalizations of future-context predictability.

	Model with backward predictability	**Model with unconditional PMI**
(Intercept)	145.981(1.85)***	145.981(1.85)***
Unigram Predictability	-23.676(0.12)***	-56.252(0.40)***
Predictability from the past	29.501(0.39)***	29.501(0.39)***
Predictability from the future	-32.576(0.40)***	-32.576(0.40)***
Word Length (in syllables)	88.569(0.38)***	88.569(0.38)***
Speech Rate	-32.767(0.14)***	-33.163(0.14)***
Speaker Age	0.070(0.04)	0.090(0.04)
Speaker Sex:M	-12.035(0.86)***	-12.035(0.86)***

Table 7: Regression coefficients from models with backward predictability and unconditional PMI as operationalizations of future context predictability. Probabilistic predictors are bolded. Parentheses denote standard error. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (ns).

Estimated effect sizes for the probabilistic predictors are presented in Table 7. Model coefficients for backward predictability and unconditional PMI are identical across the two models, with both exhibiting an expected inverse relationship with word duration ($\beta = -32.576,\ SE = 0.40,\ p < 0.001$). Similarly, unigram predictability exhibits an inverse relationship with duration in both models, though this effect is weaker in the backward predictability model ($\beta = -56.252,\ SE = 0.39,\ p < 0.001$) compared to the unconditional PMI model ($\beta = 29.501,\ SE = 0.40,\ p < 0.001$). Intriguingly, the coefficients for forward predictability were also identical across the two models but exhibited an unexpected positive effect on duration ($\beta = 29.501,\ SE = 0.40,\ p < 0.001$). This unexpected positive

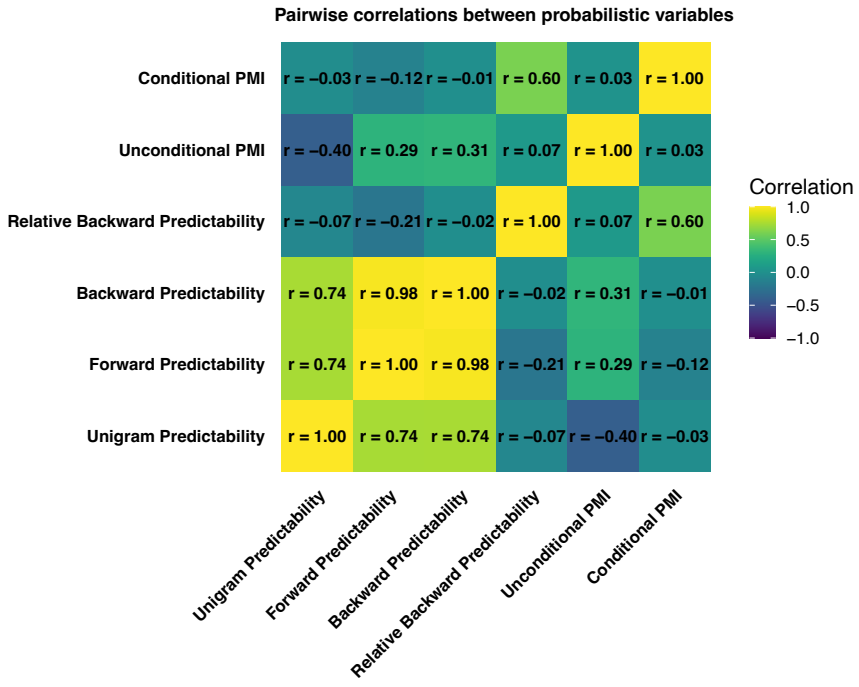


Figure 11: Pairwise Pearson correlation coefficients between all probabilistic variables for all words in the Switchboard corpus

relationship between forward predictability and duration not only contradicts previous work, but was also absent in the baseline model and in models with relative backward predictability and conditional PMI.

To determine the cause of this sign flip, we examined the pairwise correlations between all probabilistic predictors (Figure 11). First, we find that backward predictability is highly correlated with both unigram predictability ($r = 0.74$) and forward predictability ($r = 0.98$). This suggests that the observed positive effect of forward predictability likely reflects an artifact of the high pairwise correlation between these predictors. Although the model with unconditional PMI addresses the issue of collinearity between backward and unigram predictability, it does not address the issue of co-variance between forward and backward predictability, which may lead to instability in coefficient estimation.

As discussed in Section 3.2, the model with the "decorrelated" variant of backward predictability i.e., relative backward predictability not only yielded an effect size of the same magnitude as backward

predictability and unconditional PMI ($\beta = -32.576,\ SE = 0.40,\ p < 0.001$), but also produced an interpretable inverse relationship between duration and forward predictability ($\beta = -3.075,\ SE = 0.11,\ p < 0.001$). Since this measure is effectively uncorrelated with unigram predictability ($r = -0.07$) and exhibits reduced correlation with forward predictability ($r = -0.21$), we argue that relative backward predictability offers a stable alternative to both unmodified backward predictability and unconditional PMI since it mitigates issues of multi-collinearity without affecting explanatory power.

E.2 Comparison with n-gram baselines

In our main analysis, we estimate past or forward predictability and both operationalizations of future context predictability i.e., relative backward predictability and conditional PMI from the a GPT-2 language model trained for infill-based inference. Below we replicate these analysis with forward and backward transitional probabilities estimated from bigram forward and backward models with add-one smoothing.

Table 8 provides a comparison of n-gram and LM-based estimators. For both n-grams and LMs, we include both unmodified backward predictability and relative backward predictability.

	n-gram baseline	LM-estimated predictors	
		(with backward predictability)	(with relative backward predictability)
(Intercept)	135.578(1.81)***	145.981(1.85)***	145.981(1.85)***
Unigram Predictability	-23.998(0.13)***	-23.676(0.12)***	-23.675(0.12)***
Predictability from the past	2.853(0.10)	29.501(0.39)***	-3.075(0.11)***
Predictability from the future	-4.881(0.096)	-32.576(0.40)***	-32.576(0.40)***
Word Length (in syllables)	88.316(0.38)***	88.569(0.38)***	88.569(0.40)***
Speech Rate	-33.030(0.14)***	-32.766(0.14)***	-32.767(0.14)***
Speaker Age	0.077(0.04)	0.070(0.04)	0.070(0.04)
Speaker Sex:M	-11.32(0.84)	-12.034(0.86)***	-12.034(0.86)***

Table 8: Regression coefficients from models with bigram-based and LM-based predictors, with relative backward predictability as an operationalization of future context predictability. Probabilistic predictors are bolded. Parentheses denote standard error. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (ns).

Under the covariates adopted in the present study, LM-based predictors provide a significantly better fit than n-gram predictors ($\Delta\texttt{LogLik} = 1785.441, \chi^2 = 3574, p < .001$). This suggests that the relative performance of n-gram and LM-based estimators is sensitive to modeling choices. Notably, both estimators exhibit the same sign reversal for forward predictability under unmodified backward predictability, suggesting that this pattern reflects properties of the operationalization rather than the choice of estimator. In contrast, as discussed in Appendix E.1, the model with relative backward predictability yields the expected negative effect of forward predictability by reducing collinearity.

	Model with relative backward predictability	Model with conditional PMI
(Intercept)	164.13(6.40)***	145.957(6.27)***
Unigram Predictability	-23.167(0.43)***	-23.638(0.43)***
Predictability from the past	-3.091(0.42)***	-2.025 (0.41)***
Predictability from the future	-32.405(1.45)***	-38.616(2.24)***
Word Length (in syllables)	91.261(1.40)***	90.874(1.41)***
Speech Rate	-36.023(0.53)***	-36.442(0.53)***
IP-final Boundary	4.864(2.04)*	4.637(1.91)*
Speaker Age	0.0149(0.14)	0.065(0.13)
Speaker Sex:M	-10.77 (2.88)***	-9.757(2.78)***

Table 9: Regression coefficients from models with relative backward predictability and conditional PMI as operationalizations of future context predictability. Probabilistic predictors are bolded. Parentheses denote standard error. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (ns).

E.3 Replication of Study 1 with an Intonational Phrase Boundary Control

Here we replicate the analyses in Section 3 with the intonation phrase (IP)-final boundary variable. The Switchboard NXT corpus provides prosodic annotations for a smaller subset of the corpus i.e., 73 conversations. Therefore, in this replication, we only consider a subset of the original sample with Tone and Break Indices (ToBI) annotations (27,832 words). While standard prosodic controls include pitch accent and prominence, these are defined at the syllabic-level and rely on prosodic segmentation that does not align one-to-one with the word-level duration measurements provided in the annotations. Because incorporating these variables would require additional assumptions about how prominence or pitch-accent maps onto word-level units, we restrict prosodic controls to speech rate and intonational phrase boundaries.

To account for the robust intonation phrase-final lengthening effect, we extract time-aligned ToBI annotations, and create a binary intonation phrase (IP) boundary variable that encodes whether or not a word occurs at the boundary of a full intonational phrase. Specifically, only ToBI break indices of '4' and '4p' were classified as IP-final. All other break indices were classified as non-final, including 4-, which may denote annotator uncertainty between '3' (intermediate phrase boundary) and '4' (full intonation phrase boundary) (Beckman and Hirschberg, 1994).

We then fit models 1a-1c from Section 3 with IP-final phrase lengthening as an additional fixed effect. The regression coefficients for models 1a and 1b i.e., models with incrementally added relative backward and conditional PMI predictors, are presented in Table 9. Consistent with previous work, we observe a significant effect of IP phrase-final boundary in both model variants i.e., words occurring at the edge intonation phrases are more prone to lengthening. Furthermore, the effects of probabilistic

predictors on duration reported in Section 3 are qualitatively replicated when these prosodic controls are included in the model.

Moreover, replications of the incremental likelihood ratio tests with IP-final phrase boundary included yield a similar pattern of model performance. Specifically, relative backward predictability explains more variance in word duration than conditional PMI ($\Delta\texttt{LogLik} = 97.26,\ \chi^2 = 194.52, p < .001$), but a model including both predictors provides a further significant improvement in fit ($\Delta\texttt{LogLik} = 13.41,\ \chi^2 = 30.65,\ p < .001$).

We further replicated the analysis of function (N = 12760) and content (N = 13748) word durations in Section 3 with IP-final phrase boundary as a prosodic control. Regression coefficients are presented in Table 10.

	Model with relative backward predictability	Model with conditional PMI
(Intercept)	160.249(7.02)***	159.076(6.99)***
Unigram Predictability	-16.709(0.93)***	-16.9862(5.67)***
Predictability from the past	2.558(0.93)**	3.240(0.92)***
Predictability from the future	-8.273(2.12)***	-1.910(3.01)
Lexical Class:Content	4.848(0.93)	-16.304(5.67)**
IP-final Boundary	4.072(1.96)*	3.897(1.97)*
Word Length (in syllables)	83.997(1.34)***	83.447(1.35)***
Speech Rate	-29.332(0.52)***	-29.707(0.53)***
Speaker Age	0.146(0.12)	0.176(0.12)
Speaker Sex:M	-9.306(2.61)***	-8.531(2.56)**
Unigram Predictability x Lexical Class:Content	-2.741(1.03)**	-3.194(1.04)**
Predictability from the past x Lexical Class:Content	-7.377(1.03)***	-7.080(1.03)***
Predictability from the future x Lexical Class:Content	-27.206(2.82)***	-33.696(4.30)***

Table 10: Regression coefficients from models with relative backward predictability and conditional PMI as operationalizations of future context predictability. Probabilistic predictors are bolded. Parentheses denote standard error. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (ns).

Overall, the interaction between lexical class and probabilistic predictors remains qualitatively unchanged when controlling for IP-final phrase boundaries. This indicates that the main patterns reported in Section 3 are robust to the inclusion of prosodic controls.

F Study 2: Statistical Models and Results

F.1 Logistic regressions for modeling substitution choice

Table 11 provides the logistic regression models that were fit to predict substitution identity in naturalistic contexts in Section 4.

Model	Model Equation
Model 2 (Baseline)	Produced(w_t) ~ Unigram Probability + Forward Predictability + Noisy Semantic Distance + Noisy Phonetic Distance
Model 2a	Produced(w_t) ~ Unigram Probability + Forward Predictability + Noisy Semantic Distance + Noisy Phonetic Distance + **Relative Backward Predictability**
Model 2b	Produced(w_t) ~ Unigram Probability + Forward Predictability + Noisy Semantic Distance + Noisy Phonetic Distance + **Conditional PMI**
Model 2c (Both)	Produced(w_t) ~ Unigram Probability + Forward Predictability + Noisy Semantic Distance + Noisy Phonetic Distance + **Relative Backward Predictability** + **Conditional PMI**

Table 11: Logistic Regression models fit using `lme4`. Incrementally added predictors are denoted in **bold**.

F.2 Log-odds coefficients from logistic regression models

Log-odds coefficients from Models 1a and 1b are summarized in the Table 12.

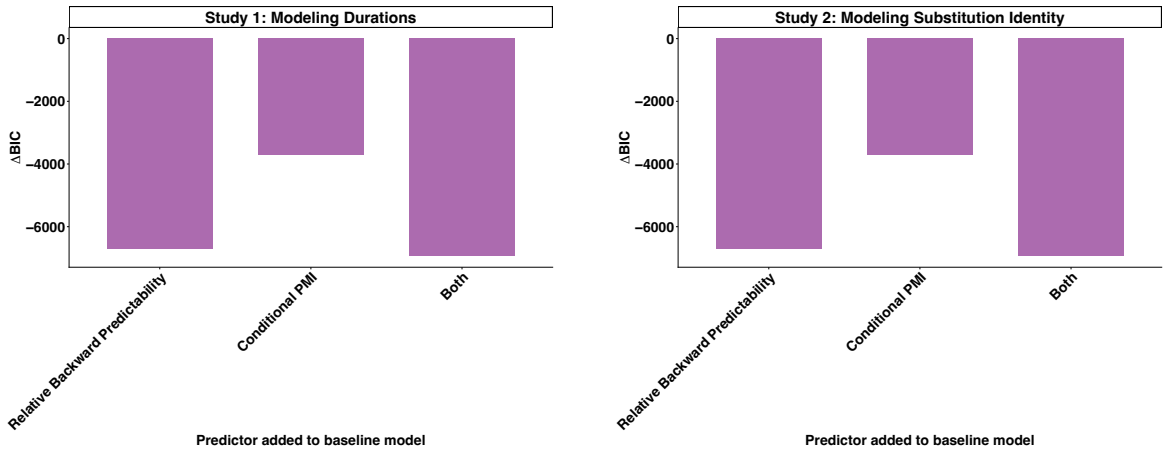
	Model 2a (with relative backward predictability)	**Model 2b** (with conditional PMI)
(Intercept)	3.523(0.19)***	3.256(0.19)***
Unigram Predictability	0.849(0.03)***	0.849(0.03)***
Predictability from the past	0.198(0.03)***	0.200(0.03)***
Predictability from the future	-0.386(0.07)***	-0.627(0.10)***
Noisy Semantic Distance	-2.400(0.19)***	-2.340(0.19)***
Noisy Phonetic Distance	-0.424(0.04)***	-0.427(0.04)***

Table 12: Log-odds coefficients from substitution models with relative backward predictability and conditional PMI as operationalizations of future context predictability. $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p > 0.05$ (ns).

G Model Comparisons with Bayesian Information Criterion

In Sections 3.2 and 4.2, we presented model comparisons to evaluate improvements in models' explanatory power when relative backward predictability and conditional PMI were added to the baseline model individually and when they were added together. Since our mixed-effects regression models had a nested structure, and because the goal of our analysis was to identify whether relative backward predictability and conditional PMI contribute redundant or unique sources of variance across both models, we used log-likelihood as the criterion for goodness of fit. To assess whether these findings hold when a more stringent measure of explanatory power is adopted, we also replicate the analysis using the Bayesian Information Criterion (BIC), which additionally penalizes the model for the number of predictors. That is, we may expect BIC to diverge from log-likelihood when comparing (i) the baseline model against models with relative backward predictability or conditional PMI added, and (ii) the model with either predictor alone against the model with both.

BIC-based comparisons for both studies are summarized in Figure 12. In Study 1, the duration model with relative backward predictability had lower BIC than the model with conditional PMI, indicating a better fit to the data ($\Delta\text{BIC} = -3012$). Adding conditional PMI to the model with relative backward predictability further reduced the BIC compared to the model with only relative backward predictability; since this difference in BIC > 10 ($\Delta\text{BIC} = -233$), we interpret it as strong evidence that the model with both predictors was a better fit to word durations, even after correcting for model complexity. In Study 2, the substitution model with conditional PMI yielded a lower BIC than the model with relative backward predictability ($\Delta\text{BIC} = -13.688$). However, when compared to a model with only conditional PMI, the model with both conditional PMI and relative backward predictability had a higher BIC, suggesting that the additional parameter did not improve model performance ($\Delta\text{BIC} = 14.514$). These results confirm that the findings about the distinct contributions of relative backward predictability and conditional PMI remain robust after adopting a more conservative, complexity-penalizing criterion for model comparison.



(a) (b)

Figure 12: Comparisons between models with (i) relative backward predictability only, (ii) conditional PMI only, and (iii) both relative backward predictability and conditional PMI. More negative ΔBIC indicates greater improvement in fit over the baseline model.

H Procedure for generating noisy phonetic target

Algorithm 2: Algorithm for random perturbation of phonetic features in target representation

Input: Target word w_t

Output: Noisy target word representation $\hat{w}_t$

$\Phi \leftarrow \{+, -, 0\}$ `// Categorical values of phonetic features`

$N \leftarrow \text{length}(w_t)$

$k \sim \mathcal{U}(1, N)$ `// Number of phonemes to select`

$p_1, \ldots, p_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}(1, N)$ `// Select positions of the` k `phonemes`

for $p \in \{p_1, \ldots, p_k\}$ **do** `// Iterate over selected` k `positions`

  $f_i \sim \mathcal{U}(1, 22)$ `// Select feature number to modify`

  $\hat{p}[f_i] \sim \mathcal{U}(\Phi \setminus p[f_i])$ `// Randomly sample an alternative feature value`

I Preprocessing of a substitution utterance with multiple errors

12. It depends on whether you whether we figure that we have a defense oriented military or an aggressive aggression oriented military

(a) Frame 1: It depends on whether you whether we figure that we have a defense oriented military or an aggression oriented military

(b) Frame 2: It depends on whether we figure that we have a defense oriented military or an aggressive aggression oriented military