UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Examining Future Context Predictability Effects in Word-form Variation and Word Choice

Permalink
https://escholarship.org/uc/item/314194w1

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Upadhye, Shiva
Futrell, Richard

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Examining Future Context Predictability Effects in Word-form Variation and Word Choice

Shiva Upadhye (shiva.upadhye@uci.edu)
Department of Language Science, University of California, Irvine
3151 Social Sciences Plaza, Irvine, CA 92697 USA

Richard Futrell (rfutrell@uci.edu)
Department of Language Science, University of California, Irvine
3151 Social Sciences Plaza, Irvine, CA 92697 USA

Abstract

Contextual predictability drives both word form and word choice in language use. The effects of the predictability of a word given its previous context are generally well-understood in both production and comprehension, but studies of naturalistic production have also revealed a poorly-understood *backward* predictability effect of a word given its *future* context, which may be related to planning. In this study, we revisit backward predictability effects using improved measures and more powerful language models, and introduce a principled measure of planning based on the pointwise mutual information between the word and the future context after controlling for the effects of previous context. We evaluate both measures for predicting word duration, and then extend the scope of these effects to a novel paradigm that involves predicting substitution errors in naturalistic productions. Our findings reveal that the proposed PMI-based measure of planning performs comparably to backward predictability. This analysis offers a useful test-bed for probing how past and future context predictability relate to underlying cognitive processes.

Keywords: Language production; Information-theoretic linguistics; Corpus Research;

Introduction

Spontaneous speech is frequently marked by *performance artifacts* that reflect how speakers navigate the cognitive demands imposed by dynamic and goal-driven communication. The prevailing view in language production research is that speakers adapt to these pressures by parallelizing the planning of upcoming context with the articulation of immediately available material (Kempen & Hoenkamp, 1987; Levelt, 1989; V. S. Ferreira & Dell, 2000; F. Ferreira & Swets, 2002; Wheeldon, Meyer, & Smith, 2006). Since these two processes are tightly yoked under this incremental strategy, one explanation for the variability observed in the speech signal is that it is symptomatic of difficulty in lexical planning.[1] (Fox Tree & Clark, 1997; Bell et al., 2003; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Gahl, Yao, & Johnson, 2012; Goldrick, Vaughn, & Murphy, 2013; Watson, Buxó-Lugo, & Simmons, 2015; Buz & Jaeger, 2016). In particular, a substantial body of work has found evidence for *probabilistic reduction* – the tendency to reduce or enhance the articulatory duration and phonetic detail of contextually predictable and unpredictable words respectively (Lieberman, 1963; Gregory, Raymond, Bell, Fosler-Lussier, & Jurafsky, 1999; Bell et al., 2003; Gahl, Garnsey, Fisher, & Matzen, 2006; Bell et al., 2009; Tily et al., 2009; Seyfarth, 2014).

Studies investigating the effects of contextual predictability on form variation have focused primarily on two probabilistic quantities: **forward** and **backward** predictability. While **forward predictability** or $p(w_t \mid C_{<t})$ denotes the probability of the current word w_t given the preceding context $C_{<t}$, **backward predictability** or $p(w_t \mid C_{>t})$ reflects its probability given an upcoming sequence $C_{>t}$. Of these two measures, forward predictability has received substantially more attention in psycholinguistic research, with a large body of work consistently revealing a robust *inverse* relationship between the facilitative effect of preceding context and various correlates of production difficulty such as word duration and disfluency (Goldman-Eisler, 1958; Jurafsky, Bell, Gregory, & Raymond, 2001a; Bell et al., 2003, 2009; Seyfarth, 2014; Dammalapati, Rajkumar, & Agarwal, 2019, 2021). Intriguingly, past work also revealed a similar inverse effect of backward predictability on duration and disfluency, after controlling for the effects of lexical frequency, forward predictability, prosodic variables, and speaker-specific characteristics (Pluymaekers, Ernestus, & Baayen, 2005; Bell et al., 2009; Seyfarth, 2014; Dammalapati et al., 2021; Chen, Levy, & Eisape, 2021). Notably, Bell et al. (2009) observed that backward predictability consistently emerged as the strongest contextual predictor of word duration in English spontaneous speech, except in the case of high-frequency function words.

Forward predictability – or (forward) surprisal – is widely accepted as a cognitively interpretable and empirically endorsed measure of incremental information processing difficulty in sentence processing (Goldman-Eisler, 1958; Ehrlich & Rayner, 1981; Jurafsky, Bell, Gregory, & Raymond, 2001b; Chang, Dell, & Bock, 2006; Hale, 2001; Levy, 2008; Futrell, Gibson, & Levy, 2020; Wilcox, Pimentel, Meister, Cotterell, & Levy, 2023; Shain, Meister, Pimentel, Cotterell, & Levy, 2024). Although understudied in comparison, backward predictability has been frequently linked to backward or look-ahead planning since it accounts for the facilitative effect of upcoming or planned context (Sinclair, 1991; Pluymaekers et al., 2005; Bell et al., 2009; Harmon & Kapatsinski, 2021) – which we refer to in this study as *future context predictability*. This construct becomes particularly relevant in on-line language production. Unlike the comprehender, a speaker has access to the conceptual content of the entire ut-

[1] Comprehensibility or audience-oriented pressures have also been shown to affect variability in word forms, see Jaeger & Buz, 2017 for a review of these positions.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

terance prior to onset of articulation (Levelt, 1989; K. Bock, Levelt, & Gernsbacher, 1994; V. S. Ferreira & Slevc, 2007). Additionally, they may also have access to lexicalized representations that may have been planned earlier but appear much later in the surface form of the utterance. For example, a planned future chunk such as "an apology for her mistake..." is more likely to constrain the choice of the current word (by strongly favoring a compatible choice such as *offered*) compared to a less informative sequence such as "an apple that she picked up..."

Despite its predictive power, backward predictability seems to imply a counter-intuitive assumption about information processing in planning: that *future* material is independent of the previously produced or *past* context [2] (henceforth, *independence* assumption; see Fig. 1). This assumption weakens the link between backward predictability and psycholinguistic theories of sentence planning: behavioral work across typologically diverse languages has found evidence against strict incrementality and in favor of flexible *out-of-order* planning that may be driven by hierarchical factors such as argument structure (Schriefers, Teruel, & Meinshausen, 1998; Lee, Brown-Schmidt, & Watson, 2013; F. Ferreira, 2013; Momma, Slevc, & Phillips, 2016; Momma & Ferreira, 2019; Nordlinger, Rodriguez, & Kidd, 2022). Furthermore, the interpretation of forward and backward predictability effects is complicated by two statistical factors. First, these are highly correlated predictors (see Bell et al., 2009), which makes it difficult to quantify the distinct predictive effects of past and future context. In addition, backward predictability is typically derived from a language model (LM) trained on reversed language input. A potential consequence of this approach is that a difference in the model may explain variance beyond the predictive effect of future context. Furthermore, existing work has tested these effects using relatively weak *n*-gram and LSTM models of predictability.

With these gaps in view, we conduct a controlled characterization of future context predictability effects in naturalistic speech. We address the methodological concerns detailed above to more precisely quantify the effects of past and future context predictability on production difficulty. Crucially, we ask if relaxing the *independence* assumption yields comparable performance to backward predictability when evaluated on symptoms of planning difficulty. To this end, we introduce an information-theoretic measure that quantifies the association between a given word and the future context under the assumption that the latter is conditioned on past context. We then conduct empirical validations of the two planning measures. In our first study, we evaluate these measures by revisiting prior work on probabilistic reduction. In the second study, we introduce a novel analysis of lexical substitutions which highlights the differential effects of past and future context predictability on word choice.

[2] We use the terms *past* and *future* context to designate their positions with respect to a word w_t in the linearized output, but make no assumptions about the time-course of planning.

Approach

As noted earlier, backward predictability (i) assumes that the current word is conditionally dependent on future context, and (ii) ignores the casual link between the past and the future context (Figure 1b). In this work, we propose a measure of future context predictability under the assumption that the speaker uses information from the past sequence to plan future production (Figure 1c).

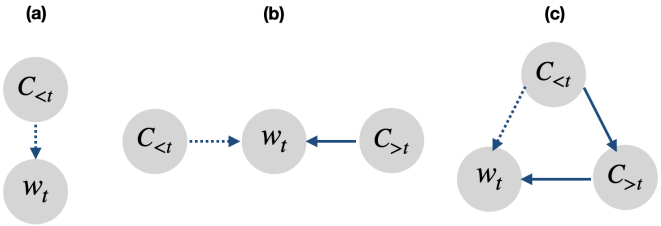

Figure 1: An illustration of the dependencies between the current word w_t, past context $C_{<t}$ and the future context $C_{>t}$ in contextual predictability measures. (a) Direct conditional dependence between w_t and $C_{<t}$ as in forward predictability; (b) w_t depends on $C_{>t}$, but no direct causal link between $C_{<t}$ and $C_{>t}$ as in backward predictability; (c) w_t depends on $C_{>t}$, and $C_{>t}$ is itself conditionally dependent on $C_{<t}$, thus relaxing the independence assumption

In particular, we define the following measure based on conditional Pointwise Mutual Information or PMI (Fano & Hawkins, 1961) of the current word (w_t) with a planned future chunk ($C_{>t}$), under the assumption that the future depends on the past ($C_{<t}$):

$$\texttt{PMI}(w_t; C_{>t} \mid C_{<t}) = \underbrace{\log p(w_t \mid C_{<t}, C_{>t})}_{\text{Bidirectional Predictability}} - \underbrace{\log p(w_t \mid C_{<t})}_{\text{Forward Predictability}} \tag{1}$$

This measure captures how informative the current word w_t is about the future context $C_{>t}$, within the past linguistic context $C_{<t}$. Thus, the above measure implicitly corrects for the effects of forward predictability. From an information-processing standpoint, this quantity reflects the planning *cost* of selecting a word w_t given the already-produced past sequence and the planned future sequence (Futrell, 2023; Upadhye & Futrell, 2022)

To conduct a principled comparison between this quantity and backward predictability, we introduce a decorrelated variant of backward predictability, defined simply as the differential between log-transformed backward and forward predictability:

$$\log p(w_t \mid C_{>t}) - \log p(w_t \mid C_{<t}) \tag{2}$$

This framing of backward predictability addresses issues of collinearity whilst also providing a direct parallel to the PMI measure, as the two quantities differ only in that PMI

assumes that the future context is planned conditional on the past, whilst decorrelated backward predictability assumes the future is planned independently of the past.

Finally, we address the potential confound that a difference in the language model may overestimate the backward predictability effect. Rather than deriving this quantity from a separate LM that learns language statistics from reversed training input, we derive it from the same LM as forward predictability (see Methods in Study 1 for further details on the language modeling approach).

Study 1: Modeling Articulatory Duration

The aim of the first study is to revisit the effects of predictability on duration found by Bell et al. (2009), with our new measures and more powerful language models.

Methods

Materials: We extract word durations from Switchboard NXT annotations (Calhoun et al., 2010; Godfrey, Holliman, & McDaniel, 1992), a word-aligned corpus of spontaneous conversations with disfluency annotations. For analysis of articulatory duration, we include only fluent utterances since disfluencies are known to strongly influence the durations of surrounding words (Fox Tree & Clark, 1997; Bell et al., 2003; Dammalapati et al., 2021).

Language Modeling: Computing forward predictability, backward predictability, and PMI requires three kinds of language models: (i) a left-to-right or forward-trained LM ($\overrightarrow{p}_\theta$), (ii) a right-to-left or backward-trained LM ($\overleftarrow{p}_\phi$), and an infill LM ($\overleftrightarrow{p}_\psi$) that computes infill predictability, respectively. In addition to the difficulty of finding an off-the-shelf large language model (LLM) capable of next-word prediction in the opposite direction, pre-trained LLMs capable of infilling differ substantially from next-word prediction models in terms of training data, objective, and model architecture. Even when trained on the same data, each model is likely to be parameterized by a different set of weights, which may capture different sources of variance in the training input. In order to address these concerns, we augment the training data to enable estimation of both forward, backward, and infill probabilities from a single GPT-2 `small` (Radford et al., 2019) LM, which we train on the CANDOR conversational speech corpus (Reece et al., 2023). Specifically, we adapt a training procedure validated by Bavarian et al. (2022), which involves randomly selecting a single word in each training sentence and transposing it to the end of the sentence (also see Donahue, Lee, & Liang, 2020). In addition, we introduce `<PRE>` and `<SUF>` tokens to demarcate the preceding and following context, and randomly shuffle their positions relative to each other. Below are examples of possible configurations for a single training sentence generated by this method:

1. Original: `<eos>` so this is the first time i did this conversation `<eos>`
2. `<eos>` `<PRE>` so this is the `<SUF>` time i did this conversation `<MID>` first `<eos>`
3. `<eos>` `<SUF>` time i did this conversation `<PRE>` so this is the `<MID>` first `<eos>`

Model hyperparameters were determined via grid search, and we selected the LM that achieved the lowest perplexity on the Switchboard corpus. During inference, the forward predictability of a word w_i was estimated as $p(w_i \mid$ `<PRE>` $C_{<t}$`<MID>`$)$, the backward predictability as $p(w_i \mid$ `<SUF>` $C_{>t}$`<MID>`$)$, and the infill probability as $p(w_i \mid$ `<PRE>` $C_{<t}$ `<SUF>` $C_{>t}$ `<MID>`$)$. We find that forward and backward predictabilities computed from this LM were significantly correlated with their counterparts estimated from separate forward-trained ($R = 0.85, p < 0.001$) and backward-trained GPT-2 models ($R = 0.83, p < 0.001$), which were also trained on CANDOR (correlations are reported between probability values estimated on the Switchboard corpus).

Statistical Analysis: We model articulatory durations using the maximally converging mixed-effects linear regression model below (Barr, Levy, Scheepers, & Tily, 2013):

`duration ~ log unigram probability + log forward predictability + word length + speech rate + speaker age + speaker gender + 1|speaker identity`

We then generated three variants of this model that differed only in terms of whether they included backward predictability, decorrelated backward predictability, or PMI as an additional predictor. All the probabilistic predictors were log-transformed. Since Bell et al. (2009) observed different effects of forward and backward predictability on the durations of function and content words, we ran two additional regressions with (i) lexical category as a main effect and (ii) interactions between lexical category and forward and backward predictability/PMI.

Results

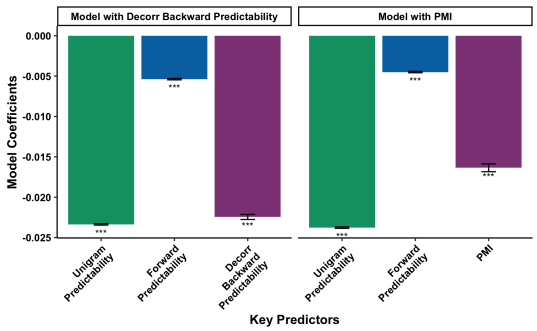

Figure 2: Coefficients of linear mixed effects models predicting articulatory word duration, for model variants with decorrelated backward predictability and PMI as planning measures. Error bars indicate standard error. $p < 0.0001(^{***})$, $p < 0.001(^{**})$, $p < 0.05(^{*})$, $p > 0.05(ns)$.

We present regression coefficients for models with decorrelated backward predictability and PMI in Figure 2. Positive coefficients predict prolongation in duration, whereas

negative coefficients predict a reduction. Across all model variants, we find that unigram predictability had a consistent and expected inverse effect on duration. Furthermore, an increase in decorrelated backward predictability ($\beta = -0.022, SE = 0.0003, p < 0.001$) and PMI ($\beta = -0.016, SE = 0.0005, p < 0.001$) both predicted reduction. Effects of forward predictability were less consistent across model variants. Whereas higher forward predictability predicted reduced duration in models with decorrelated backward predictability ($\beta = -0.00538, SE = 0.00009, p < 0.001$) and PMI ($\beta = -0.0045, SE = 0.00009, p < 0.001$), we obtain a reversal of this effect in the model with backward predictability ($\beta = 0.01706, SE = 0.0003, p < 0.001$). Upon closer inspection, forward and backward predictability were found to be highly correlated in this regression ($R = -0.96, p < 0.001$).

We conduct pairwise model comparisons to assess the improvement in explanatory power gained from adding (i) decorrelated backward predictability, (ii) PMI, and (iii) both of these planning measures to the baseline model. Our results show that the model with decorrelated backward predictability outperformed the model with PMI ($\Delta LogLik = 1991, \chi^2 = 3980.2, p < 0.001$). Furthermore, adding both backward predictability and PMI as predictors to the model led to a further increase in model fit ($LogLik = 34; \chi^2 = 68.301, p < 0.001$).

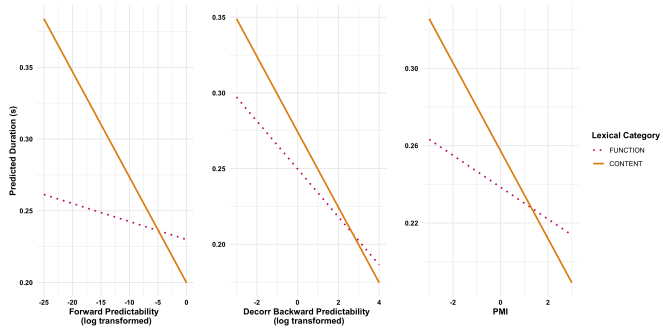

Figure 3: Regression-predicted duration effects as a function of predictability measures and the content/function word distinction. The content/function distinction modulates, but does not qualitatively change, the effects of the predictability measures.

Finally, we turn to the effect of lexical category on duration as well as the interaction between contextual predictability and lexical category. First, we replicate the expected finding that function words exhibited shorter durations than content words ($\beta = -0.0181, SE = 0.0003, p < 0.001$). Furthermore, our analysis revealed significant interactions between lexical category and forward predictability ($\beta = 0.006, SE = 0.0001, p < 0.001$), decorrelated predictability ($\beta = 0.009, SE = 0.0006, p < 0.001$), and PMI ($\beta = 0.0145, SE = 0.001, p < 0.001$). These interactions suggest that increased predictability from both past and future context leads to greater reduction in content words while function words appear less sensitive to these effects (Figure 3). The implications of these results will be discussed in more detail in the Discussion.

Study 2: Predicting Substitution Errors

The robust inverse relationship between forward and backward predictability on word durations in both prior work and the above analysis confirms that contextual predictability from both directions can reduce production difficulty. However, reduction phenomena may not be the most suitable test case for disentangling the distinct pressures imposed by past and future context predictability on word choice. For example, consider following utterances from Switchboard:

1. And I $\underbrace{\text{wasn't}}_{w_{t=3}}$ I $\underbrace{\text{didn't}}_{w^*_{t=3}}$ mean that

2. Uh you simply have to $\underbrace{\text{take}}_{w_{t=6}}$ $\underbrace{\text{accumulate}}_{w^*_{t=6}}$ your sick leave

Both utterances exemplify instances where the past context $C_{<t} = (w_1, \cdots w_{<t})$ is not particularly informative about word choice at position t. Indeed, heavily weighting forward predictability may lead to erroneous selections such as wasn't in (1) and take in (2), which are demonstrably less compatible with the future context or the speaker's overall communicative goal. Thus, a possible explanation for the self-repairs didn't and accumulate is that they not only ensure alignment with the speaker's message, but also facilitate production of the speaker's future plans.

In this study, we test this view by evaluating forward predictability, backward predictability, and PMI on a novel *generative* paradigm that predicts the lexical identity of the error in the substitution context (e.g., take in the Ex. 2).

Methods

Materials: We extract lexical substitution utterances from the Switchboard corpus, which provides *fluent*, *reparandum*, and *repair* annotations for each word in the corpus. First, we select utterances with an equal number of *reparandum* and *repair* tags to avoid cases where the speaker may have revised the structure of the utterance during repair. We then consider cases where a reparandum is followed by an immediate repair (Ex. 2), or where the reparandum sequence is repeated except for a single-word repair (Ex. 1). For utterances with multiple substitutions, we generate multiple variants such that each contains only one reparandum. Finally, we adhere to a strict definition of lexical substitutions, such that both the error (w_t) and repair/target (w^*_t) belong to the same syntactic category (Hotopf, 1980; Dell, Oppenheim, & Kittredge, 2008; Momma, Buffinton, Slevc, & Phillips, 2020).

Statistical Analysis: We aim to predict *which* word the speaker produces at a fixed position t in the substitution context. We frame this task as a binary logistic regression model. The inputs to this model are functions of a given word w_t, the

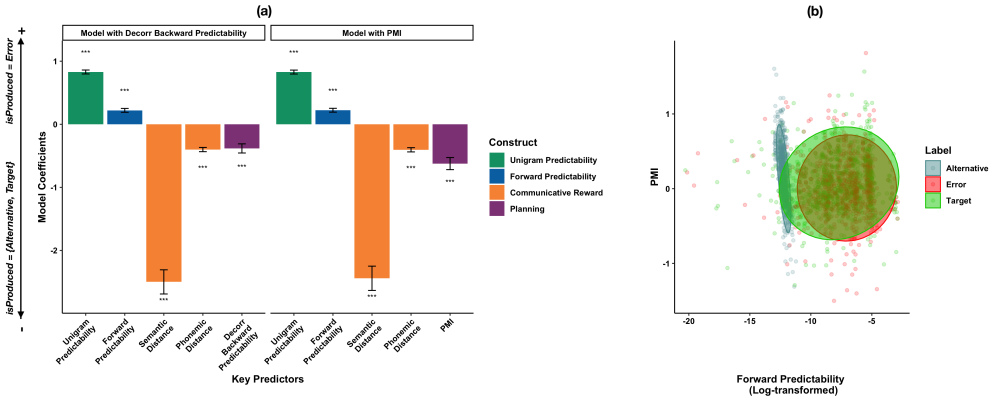

Figure 4: (a) Regression coefficients from models with decorrelated backward predictability and PMI. Error bars indicate standard error. Significance values denoted by *** for $p < 0.001$ and *ns* for $p > 0.05$; (b) Trade-off between forward predictability and PMI for errors, alternatives, and target words in substitution contexts

utterance context C, and the target word w_t^*. We include predictability measures for all words in the vocabulary, and aim to predict the identity of the observed substitution error. For example, in (2), isProduced $= 1$ when w_t = take but 0 for all other words, including the target, accumulate.

In addition to contextual predictability, we also include unigram predictability as a model predictor, since lexical frequency has been shown to affect both substitution rate and choice (Dell, 1990; Kapatsinski, 2010). Furthermore, we also include measures of alignment between a given word w_t and the speaker's goal or target word. Intuitively, words that are closer in proximity to the target in either the semantic or phonological space can serve as more rewarding or good-enough alternatives than those that are dissimilar to the target along both dimensions (V. S. Ferreira & Griffin, 2003; Goldberg & Ferreira, 2022). Thus, we operationalize the message alignment or **communicative reward** of a word w_t as a combination of its semantic and phonetic distance to the target w_t^* (Futrell, 2023). However, these standard distance measures implicitly assume that the speaker has perfect access to the target. In practice, either the target lemma or its phonological form (or both) may be inaccessible at the time of selection (Brown & McNeill, 1966; Kempen & Huijbers, 1983). We relax this assumption by generating noisy semantic and phonemic representations of the target ($\tilde{w}_t^*$). In particular, when computing semantic distance between a word and the target, we inject Gaussian noise ($\mu = 0, \sigma^2 = 0.01$) into the vectorized representation $w_t^* \in \mathbb{R}^{100}$ of the target, which we obtain from `Fasttext` embeddings (Bojanowski, Grave, Joulin, & Mikolov, 2016). For the categorical phonetic representation of the target obtained from `panphon` (Mortensen et al., 2016), we uniformly sample phonemes in the word, and for selected phoneme(s), we randomly flip a selected phonetic feature (e.g., *accumulate*: [əˈkjuːmjəˌleɪt] → [əˈɡjuːmjəˌleɪs] which reflects a change in the voicing and manner features of [k] and [t] respectively). Thus, the communicative reward of selecting w_t is defined as $\text{SemDist}(w_t, \tilde{w}_t^*) + \text{PhonDist}(w_t, \tilde{w}_t^*)$. We add all the above predictors to the following baseline model:

```
isProduced ~ unigram probability + semantic
distance + phonetic distance + forward
predictability
```

We fit two variants of this model, one with decorrelated backward predictability (Model 2a) and one with PMI (Model 2b).

Results

Model coefficients are presented in Figure 4a. A positive coefficient means that increasing the value of the predictor leads to an increase in the probability of generating the error (e.g. take in Ex. 2) whereas negative coefficient suggests a preference against producing it. In both model variants, we obtain significant positive effects of unigram predictability (2a: $\beta = 0.82927, SE = 0.03126, p < 0.001$, 2b: $\beta = 0.82927, SE = 0.03126, p < 0.001$) and forward predictability (2a : $\beta = 0.22087, SE = 0.03060, p < 0.001$, 2b : $\beta = 0.22309, SE = 0.03057, p < 0.001$), which suggests that frequency and incremental predictability may steer the speaker toward the error. In contrast, we obtain a significant negative effect of both phonetic (2a: $\beta = -0.40044, SE = 0.03352, p < 0.001$, 2b : $\beta = -0.40406, SE = 0.03358, p < 0.001$) and semantic distances (2a : $\beta = -2.49822, SE = 0.19166, p < 0.001$, 2b : $\beta = -2.44109, SE = 0.19242, p < 0.001$). This suggests that words that are dissimilar to the target are less likely to intrude or surface as errors in context. Crucially, we observe a negative effect of both decorrelated backward predictability ($\beta = -0.38231, SE = 0.07249, p < 0.001$) and PMI ($\beta = -0.62287, SE = 0.09577, p < 0.001$), contra the positive effect of both unigram and forward predictability. This suggests that an increase in future context predictability reduces the likelihood of producing the error. Similar to Study 1, we also compared the predictability of the decorrelated backward predictability and PMI to assess their explanatory power. Intriguingly, the model with PMI emerged as a better fit compared to the model with decorrelated backward predictability ($\Delta LogLik = 7$; $\chi^2 = 14.107, p < 0.001$). How-

ever, unlike Study 1, including both planning measures did not lead to an improvement in goodness-of-fit compared to the model with PMI ($\Delta LogLik = 0.9; \chi^2 = 1.6521, p > 0.05$).

Discussion

This paper presents a controlled computational-level characterization of future context predictability effects on two artifacts of production difficulty: articulatory duration and lexical substitutions. Prior work has examined backward predictability as an operationalization of this effect, which has also been linked to backward or look-ahead planning in speech (Pluymaekers et al., 2005; Bell et al., 2009; Harmon & Kapatsinski, 2021). In this work, we introduced two variants of future context predictability that were computed from the same LM and addressed issues of interpretability arising from the correlation between forward and backward predictability: (i) decorrelated backward predictability and (ii) the PMI between a current word and the speaker's future plans conditioned on past context. By assuming a causal link between the past and future context, the proposed PMI measure addresses another key concern that complicates the interpretation of backward predictability: that the speaker's planned future is independent of the previously produced context.

How do different measures of future context predictability compare? In our first study, the above planning measures were evaluated on the paradigm of modeling word durations. We found that de-correlating forward and backward predictability led to more interpretable effects whereby words are shorter when they are predictable from the past or the future. The comparison between decorrelated backward predictability and PMI, which incorporates both directions of context simultaneously, produced inconclusive results. Whereas we found that the model with decorrelated backward predictability was a better fit to word durations, PMI outperformed decorrelated backward predictability when modeling substitutions. One potential explanation for this discrepancy is that backward predictability and PMI may index distinct linear-phonological and lexico-syntactic levels of planning, respectively (K. Bock et al., 1994; V. S. Ferreira & Slevc, 2007). However, given the observational nature of this study, we tentatively conclude that the two measures produce comparable effects, with PMI providing a more interpretable account of planning since it assumes that speakers extract information from the past context and planned future during lexical planning.

Revisiting the function-content asymmetry in word durations: Prior work on modeling word durations in English observed an asymmetry in the effects of forward and backward predictability on the realization of function and content words in English: whereas durations of content words were strongly predicted by backward predictability but remained largely insensitive to the effects of forward predictability, the opposite was observed for function words (Bell et al., 2009). Our findings diverged from this result: content words were generally more prone to reduction when they were informed by past and/or future context, and these effects were attenuated but not eliminated for function words (see Ranjan, Rajkumar, and Agarwal (2022) for similar results in Hindi read-aloud speech). We speculate that this distinction could stem from (i) a difference in language models (n-gram versus representation-based) and (ii) a difference in the data used in our analysis, which excluded disfluencies.

Disentangling the effects of past and future context on word choice: A core principle of probabilistic reduction is that contextual predictability modulates the acoustic realization of the word. In our second study, we applied measures of forward and backward predictability on tasks that involved predicting the identity of the substitution error in context. Here we found that forward predictability and decorrelated backward predictability/PMI exhibited qualitatively different effects. Forward predictability along with unigram predictability—both measures which are independent of the speaker's communicative goal—predicted an increased likelihood of producing the error instead of the target. This suggests that both measures index experience-based processing or availability-based pressures in word choice (J. K. Bock, 1982; V. S. Ferreira & Dell, 2000). Therefore, the choice of substitution likely emerges from a trade-off between the counterreacting pressures of lexical availability and communicative utility (Koranda, Zettersten, & MacDonald, 2022; Futrell, 2023). However, unlike unigram and forward predictability, backward predictability and PMI correlated with a reduced likelihood of producing the substitution (see Figure 4b). We interpret this effect to mean that the substitution error may still be suboptimal from the perspective of facilitating the production of future plans. It bears mentioning, however, that the current study implicitly assumes that the speaker's future plans are deterministic (that is, it is as if the speaker knows the future context perfectly). We leave it to future work to examine how uncertainty in future plans affects these findings.

Conclusion

We have reported a critical analysis of backward predictability, a measure that indexes the facilitative effect of future context and has been likened to a future planning effect. In addition to addressing methodological concerns that complicate the interpretability of this effect vis-à-vis (forward) predictability, we also propose a theoretically-motivated PMI-based alternative that captures the informativity between a word and its future while also relaxing the assumption that future context is independent of the preceding context. We evaluate these measures on paradigms that involve predicting the articulatory form and word choice in naturalistic speech. We find that our proposed PMI-based measure produces results that are comparable to backward predictability, whilst also offering the advantage of cognitively interpretable as a measure of planning effects in speech.

References

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), 255–278.

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 preprint arXiv:2207.14255*.

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.

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. *The Journal of the acoustical society of America*, *113*(2), 1001–1024.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological review*, *89*(1), 1.

Bock, K., Levelt, W., & Gernsbacher, M. A. (1994). Language production: Grammatical encoding.

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

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, *5*(4), 325–337.

Buz, E., & Jaeger, T. F. (2016). The (in) dependence of articulation and lexical planning during isolated word production. *Language, Cognition and Neuroscience*, *31*(3), 404–424.

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.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review*, *113*(2), 234.

Chen, R., Levy, R., & Eisape, T. (2021). On factors influencing typing time: Insights from a viral online typing game. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Dammalapati, S., Rajkumar, R., & Agarwal, S. (2019, June). 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). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-3015 doi: 10.18653/v1/N19-3015

Dammalapati, S., Rajkumar, R., & Agarwal, S. (2021, February). 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. Retrieved from https://aclanthology.org/2021.scil-1.9

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and cognitive processes*, *5*(4), 313–349.

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.

Donahue, C., Lee, M., & Liang, P. (2020, July). 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. Retrieved from https://aclanthology.org/2020.acl-main.225/ doi: 10.18653/v1/2020.acl-main.225

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, *20*(6), 641–655.

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

Ferreira, F. (2013). Syntax in language production: An approach using tree-adjoining grammars. In *Aspects of language production* (pp. 303–342). 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.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, *40*(4), 296–340.

Ferreira, V. S., & Griffin, Z. M. (2003). Phonological influences on lexical (mis) selection. *Psychological Science*, *14*(1), 86–90.

Ferreira, V. S., & Slevc, L. R. (2007). Grammatical encoding. *The Oxford handbook of psycholinguistics*, 453–469.

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*(2), 151–167.

Futrell, R. (2023). Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, *120*(39), e2220593120.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*.

Gahl, S., Garnsey, S. M., Fisher, C., & Matzen, L. (2006). That sounds unlikely": Syntactic probabilities affect pronunciation. In *Proceedings of the 27th meeting of the cognitive science society.*

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce?

phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of memory and language*, *66*(4), 789–806.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on* (Vol. 1, pp. 517–520).

Goldberg, A. E., & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, *26*(4), 300–311.

Goldman-Eisler, F. (1958). The predictability of words in context and the length of pauses in speech. *Language and speech*, *1*(3), 226–231.

Goldrick, M., Vaughn, C., & Murphy, A. (2013). The effects of lexical neighbors on stop consonant articulation. *The Journal of the Acoustical Society of America*, *134*(2), EL172–EL177.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *Chicago linguistic society* (Vol. 35, pp. 151–166).

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.

Hotopf, W. (1980). Semantic similarity as a factor in whole-word slips of the tongue. *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand*, 97–109.

Jaeger, T. F., & Buz, E. (2017). Signal reduction and linguistic encoding. *The handbook of psycholinguistics*, 38–81.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001a). Evidence from reduction in lexical production. *Frequency and the emergence of linguistic structure*, *45*, 229.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001b). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, *45*, 229–254.

Kapatsinski, V. (2010). Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and speech*, *53*(1), 71–105.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive science*, *11*(2), 201–258.

Kempen, G., & Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition*, *14*(2), 185–209.

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.

Lee, E.-K., Brown-Schmidt, S., & Watson, D. G. (2013). Ways of looking ahead: Hierarchical planning in language production. *Cognition*, *129*(3), 544–562.

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

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 234–243). Honolulu, HI: Association for Computational Linguistics.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and speech*, *6*(3), 172–187.

Momma, S., Buffinton, J., Slevc, L. R., & Phillips, C. (2020). Syntactic category constrains lexical competition in speaking. *Cognition*, *197*, 104183.

Momma, S., & Ferreira, V. S. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, *114*.

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.

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 *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 3475–3484). ACL.

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.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Ranjan, S., Rajkumar, R., & Agarwal, S. (2022). Linguistic complexity and planning effects on word duration in hindi read aloud speech. *Society for Computation in Linguistics*, *5*(1).

Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., … Marin, S. (2023). The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, *9*(13), eadf3197.

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.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

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. doi:

https://doi.org/10.1073/pnas.2307876121

Sinclair, J. (1991). Corpus, concordance, collocation. *(No Title)*.

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.

Upadhye, S., & Futrell, R. (2022). Information-theoretic analysis of disfluencies in speech. In *Neurips 2022 workshop on information-theoretic principles in cognitive systems*. New Orleans. Retrieved from `https://openreview.net/pdf?id=mlUfYl4ss`$_R$

Watson, D. G., Buxó-Lugo, A., & Simmons, D. C. (2015). The effect of phonological encoding on word duration: Selection takes time. *Explicit and implicit prosody in sentence processing: Studies in honor of Janet Dean Fodor*, 85–98.

Wheeldon, L. R., Meyer, A. S., & Smith, M. (2006). Language production, incremental. *Encyclopedia of cognitive science*.

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. Retrieved from `https://aclanthology.org/2023.tacl-1.82/` doi: 10.1162/tacl$_{a0}$0612