Psychonomic Bulletin and Review (2021) 28:1365–1380
https://doi.org/10.3758/s13423-021-01886-0

BRIEF REPORT

Variable pronunciations reveal dynamic intra-speaker variation in speech planning

Oriana Kilbourn-Ceron[1] · **Matthew Goldrick**[1]

Accepted: 13 January 2021 / Published online: 26 March 2021

Abstract

In two speech production experiments, we investigated the link between phonetic variation and the scope of advance planning at the word form encoding stage. We examined cases where a word has, in addition to the pronunciation of the word in isolation, a context-specific pronunciation variant that appears only when the following word includes specific sounds. To the extent that the speaker uses the variant specific to the following context, we can infer that the phonological content of the upcoming word is included in the current planning scope. We hypothesize that the time alignment between selection of the phonetic variant in the currently-being-encoded word and retrieval of segmental details of the upcoming word is variable from moment to moment depending on current task demands and the dynamics of lexical access for each word involved. The results showed that the use of a context-sensitive phonetic variant of /t/ ("flapping") by English speakers reliably increased under conditions which favor advance planning. Our hypothesis was supported by evidence compatible with its three key predictions: an increase in flapping in phrases with a higher frequency *following* word, more flapping in a procedure with a response delay relative to a speeded response, and an attenuation of the following word frequency effect with delayed responses. This reveals that within speakers, the degree of advance planning varies continuously from moment to moment, reflecting (in part) the accessibility of form properties of individual words in the utterance.

Keywords Speech production · Psycholinguistics · Phonology

During the production of speech, there is a tension between planning multiple words in advance, which allows for fluent speech, and reducing the load on working memory, so as to alleviate interference from non-initial words (which could lead to speech errors or delays; Ferreira & Swets, 2002; Wagner et al., 2010). This paper investigates the factors which influence variation in planning scope within word form encoding processes. Previous studies have found that word form processing can take place for multiple words in advance prior to speech initiation (Oppermann et al., 2010; Schnur, 2011; Wynne et al., 2018, among others), but the degree of advance planning may differ between tasks and between individuals (Michel Lange & Laganaro, 2014; Schriefers & Teruel, 1999b). This paper advances the proposal that the extent to which multiple word forms are encoded in advance varies continuously within individuals as a function of the processing time required by each word in the utterance and of the task conditions, which can require or discourage rapid initiation of speech.

We test this hypothesis in two speech production experiments that measure the degree to which phonetic outcomes of utterance-initial words, i.e., their pronunciations, are influenced by the phonological forms of words that follow.

In our pre-registered experiments, we tested how word-specific properties (frequency and length) and task-specific demands (speeded vs. delayed productions) affected phonetic outcomes in the production of short phrases. In speeded productions, there was increased use of the contextually conditioned variant for phrases with higher-frequency words. When a response delay was imposed—allowing more planning time—there was overall more use of the contextually conditioned variant, but also a significant reduction in the effect of the second word's frequency, suggesting this frequency effect is specifically linked to advance planning of the second word. This provides new insights into the dynamic nature of advance planning during word form

✉ Oriana Kilbourn-Ceron
oriana.kilbourn-ceron@northwestern.edu

[1] Department of Linguistics, Northwestern University, 2016 Sheridan Rd., Evanston, IL 60208 USA

encoding, showing that the extent of advance planning varies not only between speakers or tasks, but from moment to moment within speakers.

Word form encoding

Word form encoding is the process of mapping a grammatical representation to its corresponding sensori-motor representation, which is used to generate speech movements. In psycholinguistic theories of speech production, word form encoding is typically assumed to require at least two stages:[1] phonological encoding and phonetic encoding. Phonological encoding involves the retrieval of the segmental content associated with selected words, construction of a prosodic frame, and association of segments to positions in the frame (see Goldrick, 2014, for a review). This frame includes minimally the syllabic level, which is organized into prosodic word groupings (a prosodic word is typically made up of a single content word plus surrounding unstressed function words; Wheeldon & Lahiri, 1997). Phonetic encoding begins once segments are associated to prosodic positions, then contextual adjustments based on syllable structure and phonological context (e.g., flapping in English) are specified in a phonetic representation (with both discrete and gradient aspects), which in turn serves as the basis for articulatory processing (see Buchwald, 2014, for a review).

Given that each word goes through several sub-stages of encoding, the question arises as to how the sub-stages of subsequent words are temporally aligned when a speaker must plan several words at once, as in typical spontaneous speech. One source of evidence comes from the contextual adjustments implemented during word form encoding, since the adjustments are often influenced by the phonology of surrounding words. For example, in many varieties of English, the final /t/ sound in the word *write* is pronounced with a shorter, voiced articulation called flap [ɾ] when it is followed by a vowel, as in *wri*[ɾ]*a letter*, but with an affricate [tʃ] in *wri*[tʃ]*you a letter* (De Jong, 1998). Since these variants are only used in those specific phonologically defined contexts (i.e., followed by a vowel or palatal glide), it must be the case that upcoming words (e.g., *a* and *you* in the above examples) have had their word forms activated sufficiently to provide a viable context for the selection of the *wri*[ɾ] or *wri*[tʃ] variants. However, the degree to which multiple word forms can or must be planned in advance is a subject of ongoing investigation.

Studies of prepared speech have shown that speakers can engage with multiple word forms prior to speech onset. The time to initiate prepared speech grows linearly with the number of phonological words, suggesting that prosodic frames for multiple words can be prepared in advance (Ferreira, 1993; Sternberg, Monsell, Knoll, & Wright, 1978; Wheeldon & Lahiri, 1997; Wheeldon & Lahiri, 2002; Wynne, Wheeldon, & Lahiri, 2018). However, these same studies also suggest that increasing the number of *syllables* only matters if they are added in the first word, suggesting that unlike prosodic words, speakers do not plan *syllabic* structure multiple words in advance (Sternberg et al., 1978; Wheeldon & Lahiri, 2002; Wynne et al., 2018). This contrast highlights the need to carefully distinguish between distinct sub-stages of word form encoding when investigating how far in advance speakers are able to plan.

Studies have also investigated whether the segmental content of non-utterance-initial words is activated prior to speech onset. Damian and Dumay (2009) used repetition priming to probe the timing of noun's phonological processing in adjective–noun phrases. Pictures described via phrases in which words shared a segment were named faster than pictures with descriptions which did not (e.g., *green goat* named faster than *green chair*). This implies that the segmental content of the second word is to some degree activated prior to speech onset. The conceptually related phonological consistency paradigm has been used to provide additional evidence that segmental content of multiple words is co-activated prior to speech onset. In this paradigm, researchers measure reaction times for determiner–adjective–noun phrases in which the determiner's pronunciation depends on the sound that follows (e.g., *a/an* in English). It has been reported that phrases with "matching" adjective and noun (e.g., *a purple giraffe*, cf. *a giraffe*) have initiation times compared to mismatching phrases like *a purple elephant* (cf. ***an*** *elephant*) (Spalek, Bock, & Schriefers, 2010). This phonological consistency effect has been replicated with determiners in other Romance and Germanic languages (Alario & Caramazza, 2002; Bürki et al., 2014, 2015, 2016; Miozzo & Caramazza, 1999). This phenomenon suggests that the specification of a determiner's form takes place at a moment when there is substantial co-activation of the segmental content of multiple subsequent words.

Studies using the picture–word interference (PWI) paradigm have shown that when naming picture displays with full sentences, speakers' initiation times are affected by distractors which are phonologically related to non-utterance-initial words (Oppermann et al., 2010; Schnur et al., 2006, 2011). Other PWI studies which elicited simple or conjoined noun phrases (e.g., *the red mouse*, or *the arrow and the bag*), have found no phonological priming

[1] It should be noted that research on speech planning is based in large part on speakers of Western European languages. As the comparative work of O'Seaghdha, Chen, and Chen (2010) demonstrates, speech planning may differ substantially for speakers of other languages in ways that are challenging to account for in current speech planning theories. The authors acknowledge that these issues limit the generalizability of the findings reviewed here, and that the same caveat should be applied to the results of the current study.

effects beyond the first prosodic word in the phrase (Meyer, 1996; Michel Lange & Laganaro, 2014; Schriefers, 1999a). However, post hoc analyses that group participants based on performance in the experiment suggest that later-occurring syllables can produce priming for participants that are relatively more accurate (Schriefers, 1999a) or respond relatively slowly (Michel Lange & Laganaro, 2014). This suggests that different possibilities are available as to how many words are phonologically activated prior to speech onset.

In contrast, studies of coarticulation in single word production have shown that speakers are capable of initiating articulation with very little prepared information (Kawamoto et al., 2015; Liu, Kawamoto, Payne, & Dorsey, 2018; Whalen, 1990). This work suggests that under certain task conditions speakers can plan for and launch articulation in the absence of phonological details about upcoming words, syllables, or even the next segment.

Studies examining spontaneous speech show that it is quite common for speakers to vary in their use of context-specific variants, even when the phonological environment remains constant (Pierrehumbert, 2001). A link has been proposed between this phenomenon and the variation in advance planning (Kilbourn-Ceron, 2017; Tamminga, MacKenzie, & Embick, 2016; Tanner, Sonderegger, & Wagner, 2017; Wagner, 2012). Kilbourn-Ceron, Clayards, and Wagner (2020) found that for a given pair of words A–B, the context-specific variant of word A is much more likely to be used when word B is predictable from word A. They argue that the predictability of word B allows it to be encoded sooner relative to word A, therefore increasing the extent of advance planning. However, the link between word-specific variables like lexical frequency and the extent of advance planning has not yet been investigated in a controlled experiment, and is the subject of the present study.

The present study

This study investigates the scope of speech planning by measuring the use of flap as a variant of a word-final /t/ in adjective-noun phrases, e.g., *great artist*, under conditions that facilitate or delay planning. Since the flap variant of /t/ only appears if a vowel follows, its presence serves as a diagnostic for simultaneous encoding of both words in the phrase. We hypothesize that the likelihood of simultaneous word form encoding for any given utterance depends on the joint influence of task demands and the planning load imposed by individual words in the utterance. This predicts that speakers should use *fewer* flaps in a task which encourages highly incremental planning, and *more* flaps when the task favors advance planning. We test this prediction across participants by comparing phrases produced in a speeded or delayed response procedure.

Our hypothesis predicts that when the word forms of nouns take longer to retrieve, vowel–initial nouns will be less likely to license the use of a flap on the preceding adjective. Our experiments test this prediction by pairing adjectives with three nouns of varying lexical frequency, a variable which is well-known to affect reaction times in picture-naming with single words (Oldfield & Wingfield, 1964; Jescheniak & Levelt, 1994), and is significantly correlated with pre-speech gaze times (Levelt & Meyer, 2000). Our hypothesis predicts that flapping will be *less* likely in a phrase with a low frequency noun, e.g., *great oyster*, compared to a phrase with a high frequency noun, e.g., *great artist*. The precise locus or loci of lexical frequency effects is still being debated, and it is likely that distinct frequency effects arise at multiple levels of processing, ranging from lexical selection and phonological encoding (Kittredge, Dell, Verkuilen, & Schwartz, 2008) to phonetic encoding (Buchwald, 2014). What is crucial for the manipulation here is that at least *some* of these word-frequency-related delays arise before phonetic encoding of the adjective is complete (and all context-related adjustments to the adjective form have been specified). Because the phonological and/or phonetic properties of lower-frequency words take longer to retrieve/specify than those of high-frequency words, they are less likely to influence phonetic encoding of the adjective.

The magnitude of the frequency effect is predicted to vary as a function of task demands. In the speeded condition, we expect that noun frequency will have a significant effect on the use of flaps, whereas as in the delayed condition the effect will be reduced, since there is more time to retrieve the noun's word form before speaking begins.

Experiment 1

Methods

The methods, design, predictions, and analysis plan for Experiment 1 were pre-registered prior to collection of data. Deviations from the pre-registered plan are noted at relevant points. The reaction time analyses were not pre-registered, but are reported for comparability with prior work. Pre-registration documents are available on this project's OSF page at https://osf.io/uge8x/.

Sample size

The target number of participants was set at 50 based on a Monte Carlo power analysis. Using effect size and variance estimates from a previous study based on spontaneous speech of the Midlands American variety (Kilbourn-Ceron et al., 2016), simulated data sets were generated, varying

the number of simulated participants (1000 simulations for each number). A mixed-effects logistic regression model was fit to each set of simulated data, and a likelihood ratio test assessed whether a significant noun frequency effect of magnitude 0.24 could be detected (non-convergent simulations were discarded without replacement). Based on these simulations, power exceeded 0.8 at 50 participants. Code and results of the power analysis are available on the OSF page.

Participants

Young adults (mean age 19.6) were recruited through the Northwestern University Linguistics department subject pool (compensated by course credit) or recruited from the Northwestern community using flyers (compensated by $7). Participants were recruited until a total of 50 met the following inclusion criteria: self-reported learning English starting before age 1, no uncorrected vision or hearing impairment, and spoke a variety of English with a productive flapping process. This last criterion was verified by the experimenter during reading of the practice items, which included non-variable flapping contexts (e.g., /t/ in *writer*, which is rarely pronounced without a flap in the target varieties of English). There were 30 female, one non-binary, and 19 male participants. Most participants reported learning English in the United States, one in Australia, and two in India. Participants' self-reported reading proficiency in English on a ten-point scale varied between 8 and 10 (mean = 9.52), and 34 reported knowing a language other than English (two declined to answer).

Materials

The phrases used for this experiment were constructed with several considerations in mind. The adjectives for the critical items were selected from a range of frequencies, and adjective frequency was included as a covariate in the statistical analysis. Assuming that the retrieval of word forms is initiated sequentially, as suggested by, e.g., Alario, Costa, and Caramazza (2002)and Levelt and Meyer (2000) , the frequency of the adjective would affect how quickly planning of the noun can begin, and potentially modulate whether noun frequency itself could have an effect. The adjectives also varied in length between one and three syllables. Previous work on dual picture naming suggests that the time alignment between initiation of articulation and gaze to the second object, which indexes planning of the second object's name, differs depending on whether the first object name is one or three syllables long (Griffin, 2003; Meyer, Belke, Häcker, & Mortensen, 2007). Meyer, Belke, Häcker, and Mortensen (2007) proposes that this is because when the first word is longer, speakers have additional time during the articulation of the first word to plan the second word. Therefore in our experiments, we might expect that longer adjectives will be more likely to use the flap variant. All nouns were two syllables long, but varied in stress pattern. Nouns with unstressed initial syllables are expected to have higher flapping rates (De Jong, 1998). Therefore, following a reviewer's suggestion, noun stress was added as a covariate in the statistical model, despite not being included in the pre-registered analysis plan.

As a final key control to isolate effects of advance planning, the phrases were constructed so as avoid existing common phrases. Previous work suggests that frequently used phrases may be stored in memory as single unit (Bybee, 2002) and therefore may have idiosyncratic phonetic realizations.

Items were prepared by selecting 40 adjectives ending in /t/ from the SUBTLEX-US database (Brysbaert & New, 2009), spanning a range of lexical frequencies (Zipf values[2] between 1.8 and 5.9, $M = 3.7$, $SD = 1.1$). The length of the adjective varied between one and three syllables (15 one-syllable, 16 two-syllable, and nine three-syllable adjectives). The one-syllable adjectives had a higher mean frequency ($M = 4.1$), but this is controlled for in the statistical analysis.

Each adjective was then paired with a unique low (Zipf value $M = 2.3$, $SD = 0.4$), medium (Zipf value $M = 3.4$, $SD = 0.3$), and high-frequency (Zipf value $M = 4.5$, $SD = 0.3$) vowel-initial noun, plus a consonant-initial noun as a distractor.

These adjective–noun bigrams were either unattested or extremely low frequency in the Google N-gram corpus (Michel et al., 2011). This yielded a total of 120 critical bigrams (three per item). Note that frequency bins were used only during item preparation, and continuous values were used for statistical analysis. The full list of items is given in Appendix A.

Design and procedure

Participants first saw ten practice trials with unrelated items. Then, adjectives were presented once per block paired with one of their four corresponding nouns. The proportion of high/medium/low frequency and consonant-initial nouns was balanced across four lists, so each block consisted of ten high-frequency nouns, ten medium frequency, ten low frequency, and ten consonant-initial, plus ten non-flapping fillers for a total of 50 trials per block. Each list

[2]The Zipf scale is equivalent to log(frequency per million words)+3, proposed by Van Heuven, Mandera, Keuleers, and Brysbaert (2014). Words on this scale are normally distributed between about 1 and 7, making it intuitively easy to compare. For reference, 1 Zipf corresponds to 0.01 frequency per million words, 3 corresponds to 1 per million, and 6 corresponds to 1000 per million.

was presented in a random order, then presented again in a random order. Within blocks, item order was also randomized. Participants were permitted to take breaks between blocks for as long as needed.

After providing informed consent, participants completed a language background questionnaire. Then, participants sat in a soundproof room at a comfortable reading distance from a computer display. Instructions and stimuli were displayed in 36pt white font on a black background. Written and verbal instructions were given to participants asking them to read aloud the phrases on the screen as quickly as possible once they appeared. Each trial began with a white fixation dot presented in the center of the screen for a randomly selected interval of 250, 500, 750, or 1000 ms. The interval was varied in order to prevent participants from falling into a repetitive, list-like intonation, which was observed during piloting (pilot participants were not included in the analysis). The stimulus was then presented for 1500 ms, and was followed by a blank screen for 500 ms. Then the next trial started. The experiment was run using the open-source software OpenSesame (Mathôt et al., 2012), and the code used to run the experiment is available on the OSF page.

Acoustic analysis

Sound files were automatically segmented into individual trials and aligned with orthographic transcriptions. Phone-level alignments were generated with the Montreal Forced Aligner (McAuliffe et al., 2017, v 1.0.0) using the English acoustic model provided with the software. The portion of each trial corresponding to an adjective-final /t/ phone was analyzed using a custom Praat script based on the method described in Eager (2015). In addition to percentage of voicing during the /t/ interval, the following acoustic measures were extracted: duration of the adjective, noun, and vowels surrounding the /t/, based on the force-aligned intervals; reaction time based on the interval between presentation of the go signal (same as stimulus onset); and the presence of a pause between the adjective and noun, determined by whether a "silence" phone was inserted by the forced alignment algorithm (minimally 30 ms due to the size of the analysis window). Figure 1 shows an example of the automatic alignment and acoustic characteristics of a flapped and non-flapped version of the same phrase from a single participant.

The dependent measure of whether or not the speaker used a flap was based on the percentage of voicing during closure, which is one of the main variables that contributes to perception of a flap (De Jong, 1998).[3] The optimal cut-off point was selected by comparing classification performance on the basis of annotations prepared by OKC. Data for 13 participants was annotated during data collection, yielding 2312 annotated tokens. In order to quantify the reliability of the percentage of voicing as an indicator of flapping, we assess the classification accuracy of three discretized versions of the voicing measure, with cut-offs at >= 50%, >= 90%, and 100%. The best overall performance comes from setting the cut-off at >= 90% voicing, which yields a balanced accuracy score of 0.88, with a sensitivity of 0.88 and specificity of 0.89. These rates are comparable to inter-annotator reliability reported in previous studies (Raymond et al., 2002). Annotations and analysis scripts are available on the OSF page.

Data exclusions

The total number of trials collected from qualifying participants was 12,000. Trials in which the participant said the wrong word, restarted, or pronounced the target phrase incorrectly were excluded ($n = 128$), as were trials in which automatic detection of reaction time or voicing failed ($n = 31$). We adopted two additional exclusion criteria which were not pre-registered. From the subset with errors removed, participants' mean response times were calculated, and responses that were ± 3 SDs away from the participants' mean were discarded ($n = 129$). Finally, we discarded trials in which the participant paused between the adjective and the noun ($n = 1731$). This is because flapping almost never occurs before a pause in English. Of the 2312 tokens annotated by hand, 238 were subsequently determined to have a pause, and only one had been perceived as a flap by the annotator. In total, these exclusions resulted in a loss of 16.82% of the data, leaving 9982 observations for analysis.

Statistical model

Flapping and reaction times were modeled mixed-effects regressions, implemented using the `lme4` (Bates et al., 2013, v. 1.1-23) package in `R` (R. Core Team, 2013, v. 1.3.959). An R Notebook detailing model specifications and outputs is available on the OSF page.

Reaction times were analyzed with a linear regression with the response variable in milliseconds, log-transformed to approach normality. Flapping was analyzed with a logistic regression, with a criterion of over 90% voicing during the aligned /t/ interval to be counted as a flap, represented with a response value of 1.

Fixed effects included adjective and noun frequency values on the Zipf scale (centered by subtracting 3.5). Since previous work suggests that length may affect the time course of inter-word phonological and articulatory

[3] While our discussion focuses on the discrete, categorical aspects of variation, there are also gradient aspects (De Jong, 1998), some of which are likely encoded within phonetic processing.

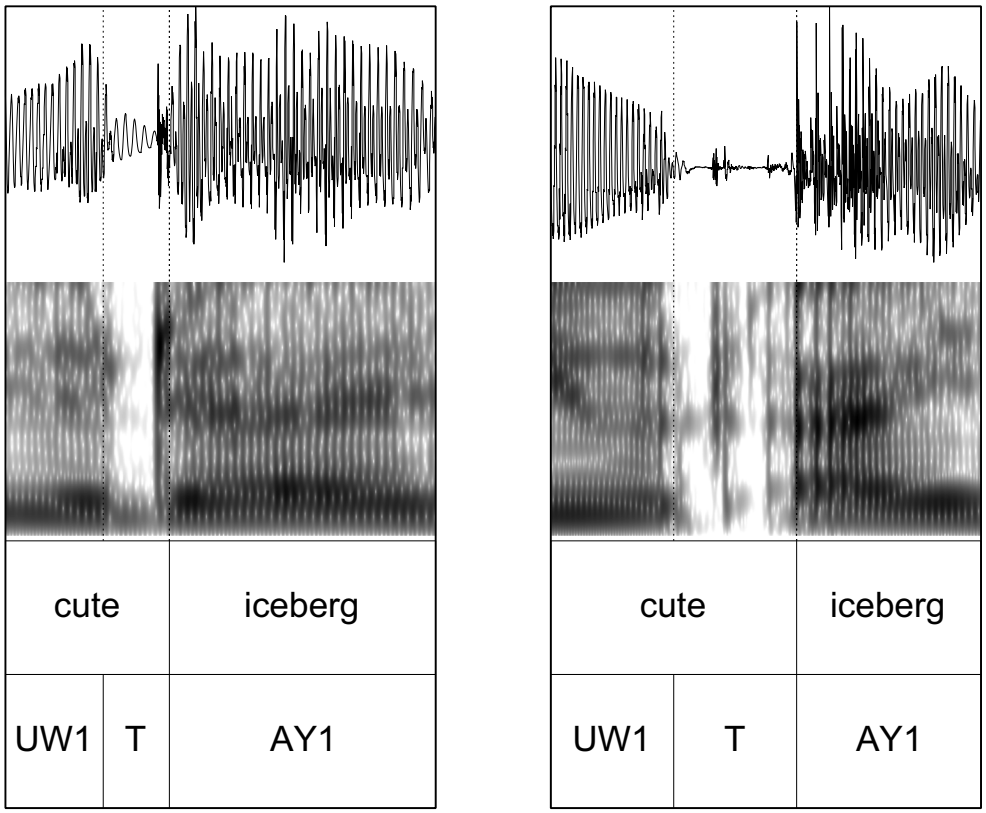

Fig. 1 Waveform (top), spectrogram (center), and automatic phone-level alignments (bottom) illustrating acoustic profiles of different possible /t/ realizations from the same speaker saying the same phrase. A flap is shown on the left, with periodic vocal fold vibration throughout, and more isolation-like glottal stop pronunciation on the right, with aperiodic vibration preceding a short silence before onset of the noun

planning (Meyer et al., 2007), we included adjective length in syllables (centered by subtracting 2). Noun stress, which had not been pre-registered, was included as a sum-coded variable, with the positive value (0.5) indicating initial stress, and the negative one (−0.5) indicating final stress. Interaction terms were also included in the model, but had not been pre-registered. Given that facilitation of adjective planning could itself allow earlier planning of the noun, we included two- and three-way interactions of length, adjective frequency, and noun frequency[4].

Additionally, block number (1 through 8, centered by subtracting 4) was included as a control variable. In the model for flapping, an additional (not pre-registered) control variable was speech rate (phones per second, excluding the interval corresponding to /t/). Speech rate was z-score normalized within speaker.

The random effects structure included random intercepts for participants and items, random slopes for noun frequency by item, and adjective frequency and noun frequency by participant. The flapping model also had a by-participant random slope for adjective length. Correlations between the random effects terms were dropped since including them in the model specification yielded a singular fit. Inclusion of random slopes for the interactions between noun frequency, adjective frequency, and noun frequency did not significantly increase goodness of fit, so they were excluded, as recommended by the selection procedure outlined in Bates, Kliegl, Vasishth, and Baayen (2015).

[4]In additional analyses, we fit a model of flapping which also included subject-level and trial-level reaction time measures (following Buz & Jaeger, 2016; Fink et al., 2018; Goldrick et al., 2019), but there were no significant effects, so reaction time predictors are excluded from the models reported below.

Results

Reaction time

Full regression results are shown in Table 1. Consistent with previous work on single-word (Griffin & Bock, 1998; Jescheniak & Levelt, 1994; Oldfield & Wingfield, 1964) and multi-word utterances (Alario, Costa, & Caramazza, 2002; Konopka, 2012), reaction times were significantly faster when the phrase-initial adjective was higher in frequency ($\hat{\beta} = -0.013$, $p < 0.001$). Replicating Alario et al. (2002), reaction times were also faster when the noun

Table 1 Fixed-effects estimates for linear regression of reaction times in Experiment 1

	β	se(β)	df	t value	Pr($\|t\|$)
(Intercept)	2.79100	0.00750	60.93	374.54	< 0.001
Adjective frequency	−0.01300	0.00260	40.25	−4.90	< 0.001
Noun frequency	−0.00530	0.00130	46.93	−4.20	< 0.001
Adjective length	0.01200	0.00340	36.17	3.40	0.002
Block number	−0.00300	0.00025	9803.61	−12.00	< 0.001
Adjective*Noun frequency	−0.00034	0.00110	34.08	−0.32	0.752
Adjective frequency*Length	−0.00690	0.00310	36.34	−2.20	0.034
Noun frequency*Adjective length	0.00100	0.00150	35.97	0.67	0.506
Adj Freq*Noun Freq*Adj Length	0.00230	0.00130	32.73	1.71	0.096

P values are estimated with Satterthwaite's degrees of freedom method from the `lmerTest` package (Kuznetsova et al., 2017). Random effects are reported on the OSF page

was higher in frequency, with the effect about half the size of the adjective frequency effect ($\hat{\beta} = -0.0053$, $p < 0.001$).

There was an increase in reaction times for phrases beginning with longer adjectives ($\hat{\beta} = 0.012$, $p = 0.002$), consistent with previous work (Sternberg et al., 1978; Wheeldon & Lahiri, 1997; Wynne et al., 2018). Length significantly interacted with adjective frequency ($\hat{\beta} = -0.0069$, $p = 0.034$), reflecting an enhancement of the frequency effect for long adjectives. No other interactions reached statistical significance ($ts < 2$).

Flapping

Full regression results are shown in Table 2. As illustrated in Fig. 2, higher-frequency nouns were significantly more likely to appear with a flap ($\hat{\beta} = 0.16$, $p < 0.001$), consistent with our prediction of a noun frequency effect. Flapping was also more likely with high-frequency adjectives ($\hat{\beta} = 0.26$, $p = 0.002$). As shown in Fig. 4 these two factors interacted ($\hat{\beta} = 0.082$, $p = 0.031$), such that noun frequency effects were strongest when the adjective was also frequent.

There was a significant effect of adjective length ($\hat{\beta} = 0.29$, $p = 0.027$), and no interactions with length were significant. As for the control covariates, flapping was more likely in later blocks ($\hat{\beta} = 0.19$, $p < 0.001$) and at faster speaking rates ($\hat{\beta} = 0.66$, $p < 0.001$), and less likely when the noun had initial stress ($\hat{\beta} = -0.94$, $p < 0.001$).

Discussion

Our examination of phonetic variation provides new evidence that the advance planning of the noun form (specifically, the initial vowel) is variable within a speaker; advance planning of the noun form is more likely as its frequency increases.

Table 2 Fixed-effects estimates for logistic regression model of flap use in Experiment 1

	β	se(β)	z	Pr($\|z\|$)
(Intercept)	−2.0000	0.210	−9.60	< 0.001
Adjective frequency	0.2600	0.086	3.07	0.002
Noun frequency	0.1600	0.045	3.63	< 0.001
Adjective length	0.2900	0.130	2.21	0.027
Block number	0.1900	0.014	13.53	< 0.001
Speech rate	0.6600	0.084	7.92	< 0.001
Noun initial stress	−0.9400	0.100	−9.10	< 0.001
Adjective*Noun frequency	0.0820	0.038	2.15	0.031
Adjective Frequency*Length	−0.1500	0.100	−1.40	0.15
Noun frequency*Adjective length	0.0091	0.055	0.17	0.868
Adj Freq*Noun Freq*Adj Length	−0.0350	0.047	−0.75	0.452

P values are estimated with Satterthwaite's degrees of freedom method from the `lmerTest` package (Kuznetsova et al., 2017). Random effects are reported on the OSF page

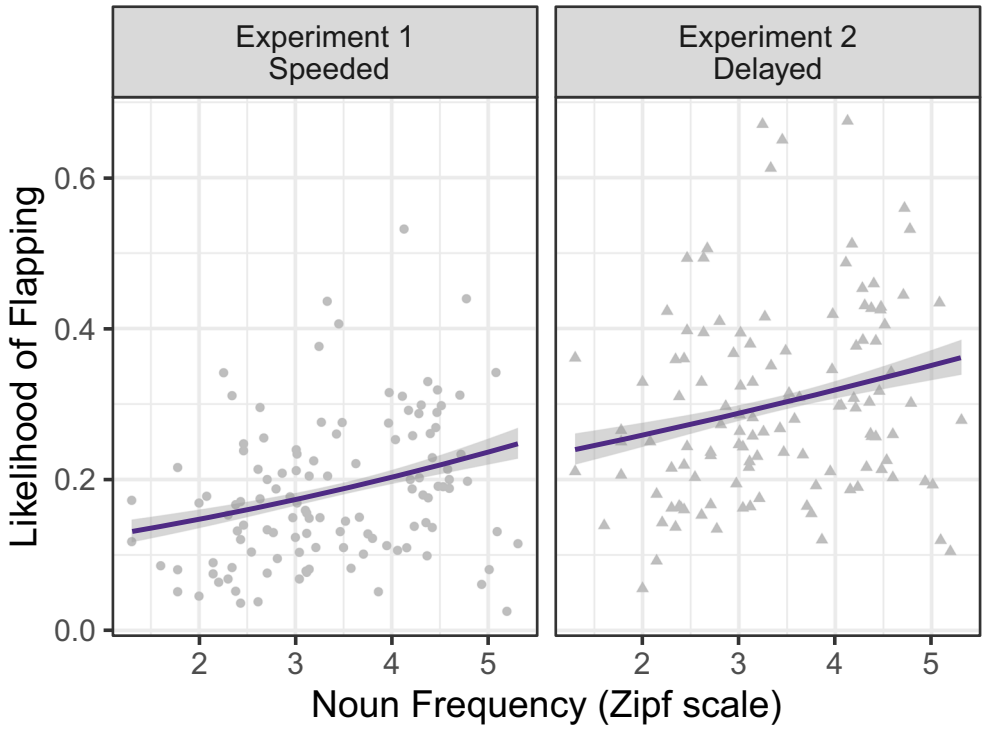

Fig. 2 Empirical plot of relationship between flapping rate (discretized) and noun frequency (Zipf scale) in Experiment 1 (speeded, online responses) and Experiment 2 (delayed responses). The line represents the estimated probability from a univariate logistic regression model and shading shows 95% confidence intervals. Each gray point represents the mean for a unique critical bigram

Adjective frequency also facilitated planning, working in concert with noun frequency. The interaction between adjective and noun frequency points towards the sequential nature of word form encoding: if noun encoding can only start once adjective encoding is complete, very low frequency adjectives can block advance planning of nouns. By contrast, more frequent adjectives will be finished planning earlier, allowing more time for noun encoding prior to speech onset.

Experiment 2

Experiment 2 was identical to Experiment 1 except that a delay was enforced between presentation of the phrase and the cue for participants to give their response. This was intended to give participants extra time to retrieve and prepare phonological details before onset of speech. Accordingly, our hypothesis predicts that flapping should be overall more likely in this condition. It also predicts that the effect of noun frequency should be reduced, since the advantage conferred by faster noun retrieval should be less important when speakers have plenty of time to retrieve the noun in advance of articulation.

Methods

An amendment to the pre-registration was made to detail changes to the participant inclusion criteria, data exclusion criteria, and model specification. The amendment is available on the OSF page.

Sample size

The target sample size was 50 participants, based on the same power analysis used for Experiment 1. However, recruitment was interrupted due to safety measures imposed by the authors' home university to mitigate the Covid-19 pandemic. Therefore, only 42 eligible participants were able to be included in this study.

Participants

Participants were recruited through the Northwestern University Linguistics department subject pool (compensated by course credit) or recruited from the Northwestern community using flyers (compensated by $7). None had participated in Experiment 1, all started learning English before age 1 and reported no uncorrected vision or hearing impair-

ment. Ages ranged from 18 to 22 ($M = 19.1$), and gender self-identifications of participants were 26 female, one non-binary, and 15 male. Participants all spoke varieties of English which include flapping (the majority learned English in the United States, one in Korea, and one in Guyana). Participants' self-reported reading proficiency in English on a ten-point scale varied between 8 and 10 ($M = 9.31$), and 34 reported knowing a language other than English.

Materials

The materials were identical to Experiment 1.

Design and procedure

The design was identical to Experiment 1, with a small change in the procedure. Written and verbal instructions were given to participants asking them to read the phrases on the screen *silently*, then say the phrase aloud as quickly as possible once the green circle appeared on the screen. Each trial began with a white fixation dot presented in the center of the screen for 500 ms, followed by presentation of the stimulus phrase. The phrase remained on the screen for a randomly selected interval of 1250, 1500, 1750, or 2000 ms. The phrase was then masked by six black Xs with a large green circle above them. This cue for participants to initiate their response stayed on the screen for 600 ms, followed by a black screen for 900 ms before the beginning of the next trial. The experiment was run using the open-source software OpenSesame (Mathôt et al., 2012), and the code used to run the experiment is available in the OSF page.

Data exclusions

The total number of trials collected from qualifying participants was 10,080. Trials in which the participant said the wrong word, restarted, or pronounced the target phrase incorrectly were excluded ($n = 80$), as were trials in which automatic detection of reaction time or voicing failed (typically due to participants speaking on or before the response prompt, $n = 69$). From the subset with errors removed, participants' mean response times were calculated, and responses that were ± 3 SDs away from the participants' mean were discarded ($n = 126$). As specified in the amendment to the pre-registration (registered prior to data collection for Experiment 2), trials in which the participant paused between the adjective and the noun were discarded ($n = 786$). In total, these exclusions resulted in a loss of 10.71% of the data, leaving 9,000 observations for analysis.

Statistical model

Models for reaction time and flapping were fit with the same specifications as in Experiment 1, reported in Tables 3 and 4. An additional cross-experiment analysis, which was not pre-registered, was conducted for flapping to test the reliability of effect size differences between the two experiments. The same model specification was used as for previous flapping models, with the addition of a "condition" variable which was set to 0 for speeded responses (Experiment 1) and to 1 for the delayed responses (Experiment 2). This variable was allowed to interact with each of the other fixed effects. The full model is reported

Table 3 Fixed-effects estimates for linear regression of reaction times in Experiment 2 (delayed responses)

	β	se(β)	df	t value	Pr($\|t\|$)
(Intercept)	2.59400	0.00950	49.22	273.30	< 0.001
Adjective frequency	−0.00250	0.00310	43.78	−0.78	0.438
Noun frequency	−0.00270	0.00130	18.36	−2.20	0.043
Adjective length	0.00910	0.00410	36.21	2.22	0.032
Block number	−0.00260	0.00043	8822.48	−6.00	< 0.001
Adjective*Noun frequency	0.00075	0.00100	14.02	0.72	0.483
Adjective frequency*Length	−0.00620	0.00370	36.40	−1.70	0.106
Noun frequency*Adjective length	0.00180	0.00150	16.29	1.21	0.244
Adj Freq*Noun Freq*Adj Length	0.00140	0.00130	11.20	1.08	0.302

P values are estimated with Satterthwaite's degrees of freedom method from the `lmerTest` package (Kuznetsova et al., 2017). Random effects are reported on the OSF page

Table 4 Fixed-effects estimates for logistic regression model of flap use in Experiment 2 (delayed responses)

	β	se(β)	z	Pr(|z|)
(Intercept)	−1.200	0.230	−5.00	< 0.001
Adjective frequency	0.250	0.092	2.71	0.007
Noun frequency	0.058	0.050	1.16	0.247
Adjective length	0.210	0.140	1.52	0.129
Block number	0.100	0.012	8.47	< 0.001
Speech rate	0.740	0.080	9.17	< 0.001
Noun initial stress	−0.760	0.100	−7.40	< 0.001
Adjective*Noun frequency	0.059	0.042	1.41	0.16
Adjective frequency*Length	−0.120	0.110	−1.10	0.268
Noun frequency*Adjective length	−0.015	0.060	−0.26	0.796
Adj Freq*Noun Freq*Adj Length	−0.140	0.052	−2.60	0.008

P values are estimated with Satterthwaite's degrees of freedom method from the `lmerTest` package (Kuznetsova et al., 2017). Random effects are reported on the OSF page

in Appendix B. Given that the power analysis was not conducted with "condition" or any of its interactions in mind, the results of this pooled Experiment 1/2 model should be taken as exploratory rather than confirmatory, and should be confirmed through future replications.

Results

Reaction time

Consistent with more advance planning, mean reaction times were faster in Experiment 2 (means for Experiment 1: 621 ms; Experiment 2: 403 ms). Reaction times were faster for higher frequency nouns ($\hat{\beta} = -0.0027$, $p = 0.043$), but in contrast to Experiment 1 there was no significant effect of adjective frequency ($\hat{\beta} = -0.0025$, $p = 0.438$).

There was a significant effect of adjective length in Experiment 2 ($\hat{\beta} = 0.0091$, $p = 0.032$). Unlike Experiment 1, there was no significant interaction between adjective length and frequency ($\hat{\beta} = -0.0062$, $p = 0.106$), nor were any other interactions significant.

Flapping

As predicted, there was significantly more flap use in Experiment 2 (mean 29.9%), the delayed response condition, as compared to Experiment 1 (mean 18.6%; cross-experiment model, delayed condition: $\hat{\beta} = 0.91$, $p = 0.001$).

As in Experiment 1, higher frequency adjectives were more likely to be flapped ($\hat{\beta} = 0.25$, $p = 0.007$). Figure 2 (right panel) suggests that noun frequency was also associated with more flap use, but this effect was not significant ($\hat{\beta} = 0.058$, $p = 0.247$). Consistent with the qualitative difference in noun frequency effects across experiments, the pooled model finds a significant interaction of condition and noun frequency effects, illustrated in Fig. 3 ($\hat{\beta} = -0.11$, $p = 0.044$) This suggests that noun frequency effects on flapping are driven, in part, by advance planning.

The interaction between adjective and noun frequency was not significant, in contrast to Experiment 1 ($\hat{\beta} = 0.059$, $p = 0.16$; see Fig. 4). There was a significant three-way interaction between adjective frequency, adjective length, and noun frequency ($\hat{\beta} = -0.14$, $p = 0.008$), such that the adjective frequency * noun frequency interaction observed in Experiment 1 was primarily found for monosyllabic adjectives. No other interactions were significant.

The covariate effects were qualitatively similar to Experiment 1, with positive effects for block number and speech rate (Block Number: $\hat{\beta} = 0.1$, $p < 0.001$, speech rate: $\hat{\beta} = 0.74$, $p < 0.001$), and a negative effect for initial stress on the noun ($\hat{\beta} = -0.76$, $p < 0.001$).

Discussion

As predicted, allowing more time for planning decreased reaction times and increased the use of flapping. Critically, the effect of noun frequency was significantly attenuated with delayed responses. This suggests that noun frequency effects on flapping are driven in part by an advance planning process that varies depending on task demands.

General discussion

This study investigated the within-speaker dynamics of word form encoding in multi-word utterances. We focused

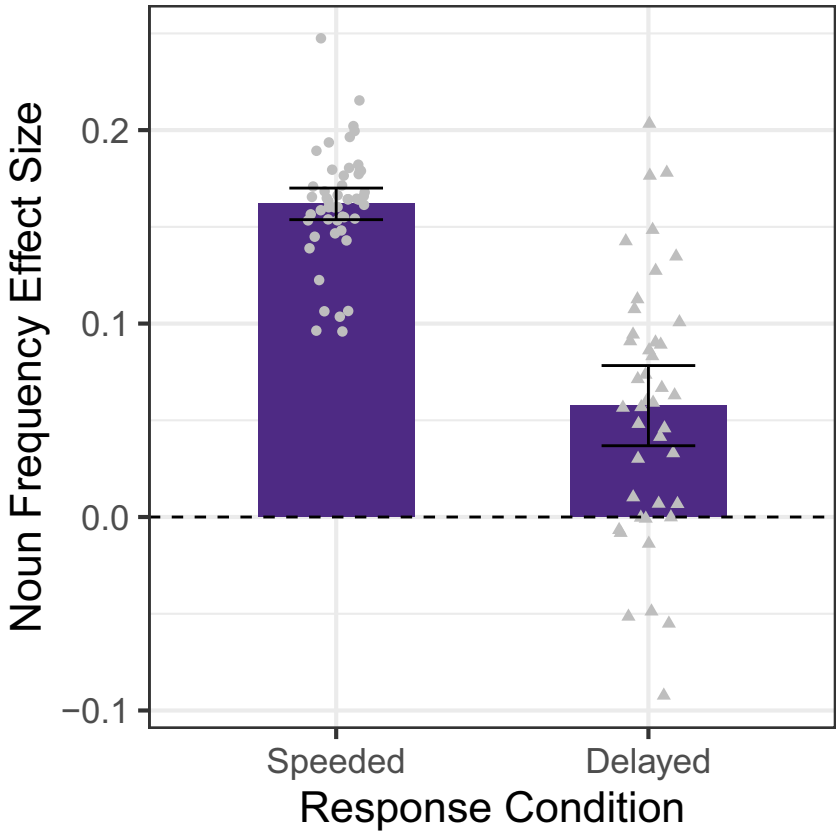

Fig. 3 Comparison of noun frequency effect size in Experiment 1 (speeded) and Experiment 2 (delayed). Bar height represents the estimated fixed effect size, and error bars represent 95% confidence intervals based on the subject-level variance of random slope for noun frequency. Individual gray points show subject random slope estimates

on a probabilistic phonological pattern in which there is a dependency between two adjacent phonemes belonging to different words, namely /t/-flapping in English. Since the use of a flap variant requires "look ahead" to check whether the next word begins with a vowel, the presence of a flap serves as an index advance planning. Our results show that in adjective-noun phrases, the probability of flap use, and therefore the degree of advance planning, is based on word-specific utterance characteristics (lexical frequency) in addition to current task demands. Flapping was more likely to occur when nouns were easier to retrieve (i.e., higher frequency). When a response delay was enforced, more advance planning occurred, diminishing the disadvantage of low frequency nouns and increasing the overall likelihood of flap use.

These findings converge with previous work showing that advance planning can shift as a function of task demands (Griffin, 2003; Klaus et al., 2017; Meyer et al., 2007; Wagner, Jescheniak, & Schriefers, 2010; Wynne et al., 2018). Our results complement previous work showing that, under the same demands, different speakers may show different degrees of advance planning (Michel Lange & Laganaro, 2014; Schriefers & Teruel, 1999b). This study adds a new key insight to the general concept of flexible planning scope: within speakers, the degree of advance planning varies continuously from moment to moment, partly as a function of the accessibility of the form of upcoming words (as indexed by lexical frequency).

Our results also converge with work on phonetic variation in spontaneous speech, supporting the causal link between advance planning and variation proposed in Kilbourn-Ceron (2017). Kilbourn-Ceron et al. (2020) investigated flapping in spontaneous speech, and found that higher conditional probability of the upcoming word given the target word (e.g., the probability of *artist* coming after *great*) led to increased likelihood of flapping. They did not find any effect of second word frequency, as we did in this study. However, the two measures are highly correlated in spontaneous speech, making it difficult to disentangle their effects. Future work could investigate the effect of conditional probability experimentally, where these two factors can be de-correlated. Our proposal predicts that there should indeed be an effect proportional to the influence of conditional probability on the degree of advance planning. Some preliminary supporting evidence has been found for liaison in French (Wagner, Lachapelle, and Kilbourn-Ceron, 2020)

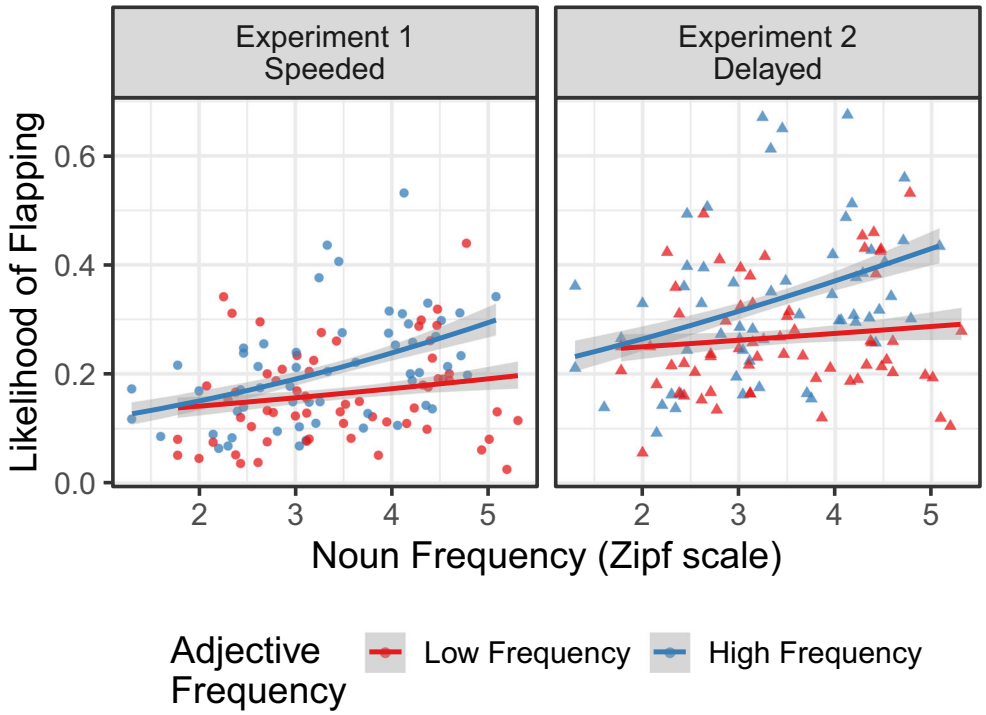

Fig. 4 Empirical plot of relationship between flapping rate (discretized) and noun frequency (Zipf scale) in Experiment 1 (speeded, online responses) and Experiment 2 (delayed responses). Panels show data split into upper and lower quantiles by adjective frequency. The line represents the estimated probability from a univariate logistic regression model and shading shows 95% confidence intervals. Each point represents the mean for a unique critical bigram, colored by adjective frequency quantile

The manipulations in this study targeted individual words (frequency, length) and global task demands (response delays). It is likely that many other factors could affect the degree of advance planning at the word form stage. Even within speech planning, the advance planning of word forms must be bounded by the extent of advance planning at earlier stages, at least according to serial models of speech planning. Future work should investigate whether delays in processing of semantic and grammatical aspects of the utterance have downstream consequences for the extent of advance planning at the word form level.

Conclusions

This paper provides new evidence for the dynamic nature of advance planning during word form encoding. Phonetic variation provides us with a new tool to investigate the scope of planning, moving beyond reaction time to examine the ongoing nature of planning following the onset of speech.

Appendix A: Experimental items

Table 5 List of materials used in Experiments 1 and 2

	Noun Frequency			
Adjective	High	Medium	Low	(Consonant)
animate	ocean	oatmeal	acorn	soil
erudite	office	invoice	armchairs	map
starlit	opera	autumn	expanse	ridge
tacit	arrest	exile	accords	retreat
russet	island	apron	okra	cellar
ornate	army	outpost	archway	ransom
literate	owner	ethic	outcry	rumor
inert	agent	error	inbox	lemur

Table 5 (continued)

Adjective	Noun Frequency			
	High	Medium	Low	(Consonant)
fraught	estate	illness	upkeep	mask
taut	order	eyebrow	airstrike	mummy
trite	issue	export	anthems	moral
petite	event	oven	icons	shorts
upbeat	attempt	outlaw	envoys	shepherd
moot	effect	absence	allure	roads
sedate	adult	insect	oboe	knight
distraught	affair	outing	aardvark	lighter
mute	uncle	outcast	embers	niece
adequate	island	eggnog	oxfords	lipstick
curt	exchange	athlete	emcee	ranger
scarlet	outfit	icing	anvil	robe
elaborate	alarm	archive	acclaim	method
concrete	address	abyss	alcove	ceiling
delicate	apple	onion	aloe	sausage
slight	effort	impact	algae	menu
corporate	ally	orchid	android	relative
neat	eagle	almond	adverb	rider
polite	action	ogre	orca	maniac
remote	offense	orbit	armoire	shield
favorite	artist	oyster	easel	saddle
opposite	interest	ointment	eyedrop	yacht
wet	airplane	ostrich	urchins	stream
bright	evening	opal	orchards	necklace
complete	exit	altar	antler	madness
private	angle	eclipse	igloo	web
fat	actor	eggplant	earthworm	shrimp
cute	airport	iceberg	aussie	monkey
quiet	award	exhaust	alehouse	raid
sweet	alley	amber	anklet	rash
late	image	intake	ingot	symbol
great	exam	atlas	airwave	spoon

Each row represents one item, showing the adjective and three unique nouns that were paired with it. For full details including frequency measures, visit this project's OSF page

Appendix B: Logistic model comparing experiments

Table 6 Fixed-effects estimates for logistic mixed model of pooled data from Experiments 1 and 2

	β	se(β)	z	Pr(|z|)
(Intercept)	−2.1000	0.210	−9.700	< 0.001
Condition (Delayed)	0.9100	0.290	3.190	0.001
Adjective Frequency	0.2500	0.086	2.860	0.004
Noun Frequency	0.1700	0.054	3.090	0.002
Adjective Length	0.2000	0.120	1.660	0.098
Block Number	0.1800	0.014	13.440	< 0.001
Speech Rate	0.7600	0.078	9.690	< 0.001
Noun Initial Stress	−0.8000	0.095	−8.400	< 0.001
Condition*Adjective Frequency	−0.0010	0.050	−0.020	0.984
Condition*Noun Frequency	−0.1100	0.054	−2.000	0.044
Adjective*Noun Frequency	0.0900	0.046	1.970	0.049
Condition*Adjective Length	−0.0041	0.064	−0.064	0.949
Adjective Frequency*Length	−0.1500	0.110	−1.500	0.144
Noun Frequency*Adjective Length	0.0018	0.066	0.027	0.978
Condition*Speech Rate	−0.0190	0.096	−0.200	0.84
Condition*Block Number	−0.0860	0.018	−4.700	< 0.001
Condition*Noun Initial Stress	0.0300	0.100	0.290	0.775
Condition*Adjective Freq*Noun Freq	−0.0220	0.042	−0.530	0.597
Condition*Adjective Freq* Adjective Length	0.0320	0.054	0.590	0.552
Condition*Noun Freq* Adjective Length	−0.0120	0.062	−0.190	0.85
Adj Freq*Noun Freq*Adj Length	−0.0310	0.057	−0.530	0.594
Condition*Adj Freq*Noun Freq*Adj Length	−0.1000	0.051	−2.000	0.049

P values are estimated with Satterthwaite's degrees of freedom method from the `lmerTest` package (Kuznetsova et al., 2017). Random effects are reported on the OSF page

Acknowledgements We wish to acknowledge the contributions of Chandana Sooranhalli in running participants, and Chun Liang Chan for technical assistance with software and equipment. We are also grateful to the participants of the Phonatics discussion group at Northwestern and the audience of LabPhon17 for questions and comments that improved this work. We also wish to thank. This research was supported by the Fonds de Recherche du Québec through Postdoctoral fellowship 2019-B3Z-255232 awarded to OKC.

Author Contributions (following CRediT: https://casrai.org/credit/): OKC: Conceptualization, Data Curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. MG: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Funding sources Fonds de recherche du Québec: Société et culture

Open Practices A detailed analysis plan including methods, sample size, and statistical model specification was pre-registered prior data collection for Experiment 1. An amendment to the plan was pre-registered prior to data collection for Experiment 2. These documents are available in this project's Open Science Framework page, along with code for the experimental procedure, code and results of the power analysis, code for the statistical analysis, dependent measure data, and acoustic data (for those participants who consented to making their data publicly available). The OSF page can be found at https://osf.io/uge8x/.

References

Alario, F.-X., & Caramazza, A. (2002). The production of determiners: Evidence from French. *Cognition*, *82*(3), 179–223.

Alario, F.-X., Costa, A., & Caramazza, A. (2002). Frequency effects in noun phrase production: Implications for models of lexical access. *Language and Cognitive Processes*, *17*(3), 299–319.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4 [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 (R package version 1.1-21).

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. arXiv:1506.04967

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Buchwald, A. (2014). Phonetic processing. In Goldrick, M., Ferreira, V., & Miozzo, M. (Eds.) *The Oxford handbook of language production*, (pp. 245–258). Oxford: Oxford University Press.

Bürki, A., Laganaro, M., & Alario, F.-X. (2014). Phonologically driven variability: The case of determiners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1348–1362.

Bürki, A., Frauenfelder, U. H., & Alario, F.-X. (2015). On the resolution of phonological constraints in spoken production: Acoustic and response time evidence. *The Journal of the Acoustical Society of America*, *138*(4), EL429-EL434. https://doi.org/10.1121/1.4934179

Bürki, A., Sadat, J., Dubarry, A.-S., & Alario, F.-X. (2016). Sequential processing during noun phrase production. *Cognition*, *146*, 90–99. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027715300585. https://doi.org/10.1016/j.cognition.2015.09.002

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.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition*, 215–221.

Damian, M. F., & Dumay, N. (2009). Exploring phonological encoding through repeated segments. *Language and Cognitive Processes*, *24*(5), 685–712.

De Jong, K. (1998). Stress-related variation in the articulation of coda alveolar stops: Flapping revisited. *Journal of Phonetics*, *26*(3), 283–310.

Eager, C. (2015). Automated voicing analysis in Praat: Statistically equivalent to manual segmentation. In The Scottish Consortium for ICPhS 2015 (Ed.) *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: University of Glasgow.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review*, *100*(2), 233–253.

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*, 57–84.

Fink, A., Oppenheim, G. M., & Goldrick, M. (2018). Interactions between lexical access and articulation. *Language, Cognition and Neuroscience*, *33*(1), 12–24.

Goldrick, M. (2014). Phonological processing: The retrieval and encoding of word form information in speech production. In Goldrick, M., Ferreira, V., & Miozzo, M. (Eds.) *The Oxford handbook of language production*, (pp. 228–244). Oxford: Oxford University Press.

Goldrick, M., McClain, R., Cibelli, E., Adi, Y., Gustafson, E., Moers, C., & Keshet, J. (2019). The influence of lexical selection disruptions on articulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1107.

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, *38*, 313–338.

Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychonomic Bulletin & Review*, *10*(3), 603–609.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824.

Kawamoto, A. H., Liu, Q., & Kello, C. T. (2015). The segment as the minimal planning unit in speech production and reading aloud: evidence and implications. *Frontiers in Psychology*, *6*, 1457. Retrieved from https://doi.org/10.3389/fpsyg.2015.01457

Kilbourn-Ceron, O., Wagner, M., & Clayards, M. (2016). The effect of production planning locality on external sandhi: A study in /t/. In *The proceedings of the 52nd Meeting of the Chicago Linguistics Society*. Retrieved from http://ling.auf.net/lingbuzz/003119

Kilbourn-Ceron, O. (2017). *Speech production planning affects variation in external sandhi (Unpublished doctoral dissertation)*. McGill University.

Kilbourn-Ceron, O., Clayards, M., & Wagner, M. (2020). Predictability modulates pronunciation variants through speech planning effects: A case study on coronal stop realizations. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *11*(1), 5. https://doi.org/10.5334/labphon.168

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.

Klaus, J., Mädebach, A., Oppermann, F., & Jescheniak, J. D. (2017). Planning sentences while doing other things at the same time: Effects of concurrent verbal and visuospatial working memory load. *Quarterly Journal of Experimental Psychology*, *70*(4), 811–831.

Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language*, *66*(1), 143–162.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Levelt, W. J., & Meyer, A. S. (2000). Word for word: Multiple lexical access in speech production. *European Journal of Cognitive Psychology*, *12*(4), 433–452.

Liu, Q., Kawamoto, A. H., Payne, K. K., & Dorsey, G. N. (2018). Anticipatory coarticulation and the minimal planning unit of speech. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(1), 139.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. Retrieved from https://doi.org/10.3758/s13428-011-0168-7.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th conference of the International Speech Communication Association* (pp. 498–502).

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

Meyer, A. S., Belke, E., Häcker, C., & Mortensen, L. (2007). Use of word length information in utterance planning. *Journal of Memory and Language*, *57*(2), 210–231. https://doi.org/10.1016/j.jml.2006.10.005

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., & Pickett, J. P. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Michel Lange, V., & Laganaro, M. (2014). Inter-subject variability modulates phonological advance planning in the production of adjective-noun phrases. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.00043

Miozzo, M., & Caramazza, A. (1999). The selection of determiners in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 907.

Oldfield, R. C., & Wingfield, A. (1964). The time it takes to name an object. *Nature*, *202*, 1031–1032.

Oppermann, F., Jescheniak, J. D., & Schriefers, H. (2010). Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X10000628, https://doi.org/10.1016/j.jml.2010.07.004. *Journal of Memory and Language*, *63*(4), 526–540.

O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, *115*(2), 282–302.

Pierrehumbert, J. B. (2001). Stochastic phonology. *Glot International*, *5*(6), 195–207.

R. Core Team (2013). A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raymond, W. D., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R., & Hilts, C. (2002). An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. In *Seventh International Conference on Spoken Language Processing*.

Schnur, T. T., Costa, A., & Caramazza, A. (2006). Planning at the phonological level during sentence production. *Journal of Psycholinguistic Research*, *35*(2), 189–213. https://doi.org/10.1007/s10936-005-9011-6

Schnur, T. T. (2011). Phonological planning during sentence production: Beyond the verb. *Frontiers in Psychology, 2*. Retrieved from https://doi.org/10.3389/fpsyg.2011.00319

Schriefers, H. (1999a). Phonological facilitation in the production of two-word utterances. *European Journal of Cognitive Psychology*, *11*(1), 17–50.

Schriefers, H., & Teruel, E. (1999b). The production of noun phrases: A cross linguistic comparison of French and German. In *Proceedings of the 21st annual conference of the Cognitive Science Society*, (pp. 637–642). Mahwah: Erlbaum.

Spalek, K., Bock, K., & Schriefers, H. (2010). A purple giraffe is faster than a purple elephant: Inconsistent phonology affects determiner selection in English. *Cognition*, *114*(1), 123–128. https://doi.org/10.1016/j.cognition.2009.09.011

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In Stelmach, G. (Ed.) *Information processing in motor control and learning*, (pp. 117—152). New York: Academic Press.

Tamminga, M., MacKenzie, L., & Embick, D. (2016). The dynamics of variation in individuals. *Linguistic Variation*, *16*(2), 300–336.

Tanner, J., Sonderegger, M., & Wagner, M. (2017). Production planning and coronal stop deletion in spontaneous speech. *Laboratory Phonology, 8*, 1.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176–1190.

Wagner, M. (2012). Locality in phonology and production planning. In Loughran, J., & McKillen, A. (Eds.) *Proceedings of phonology in the 21 century: Papers in honour of Glyne Piggott*, (Vol. 22, pp. 1–18). Montreal.

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.

Wagner, M., Lachapelle, J., & Kilbourn-Ceron, O. (2020). *Liaison and the locality of production planning*. Poster presented at the 17th Conference on Laboratory Phonology, July 6-8. [Online].

Whalen, D. (1990). Coarticulation is largely planned. *Journal of Phonetics*, *18*, 3–35.

Wheeldon, L. R., & Lahiri, A. (1997). Prosodic units in speech production. *Journal of Memory and Language*, *37*, 356–381.

Wheeldon, L. R., & Lahiri, A. (2002). The minimal unit of phonological encoding: Prosodic or lexical word. *Cognition*, *85*(2), 31–41.

Wynne, H. S., Wheeldon, L., & Lahiri, A. (2018). Compounds, phrases and clitics in connected speech. *Journal of Memory and Language*, *98*, 45–58.