Running Head: FLEXIBLE PERCEPTUAL SENSITIVITY

Flexible perceptual sensitivity to acoustic and distributional cues

Clara Cohen[1,3], Shinae Kang[2,3]

[1] University of Glasgow; [2]The George Washington University; [3]University of California, Berkeley

Address for correspondence: Clara Cohen, English Language and Linguistics, 12 University Gardens, Glasgow G12 8QQ, UK. Email: clara.cohen@glasgow.ac.uk

Author note: This work was completed with funding by the University of California, Berkeley's Linguistics Research Apprenticeship Program and the University of Glasgow College of Arts Strategic Research Allocation Fund. We are grateful to Rozina Fonyo for her vital assistance in preparing and running experiments, and to Keith Johnson for helpful discussion and suggestions. We are also grateful to Professor James Mahshie for supporting data collection with Experiment 1b. Portions of this work have been presented at the 2015 CUNY Conference on Sentence Processing, the Fall 2015 Acoustical Society of America Meeting, and the 2016 Annual Meeting of the Berkeley Linguistics Society.

Abstract

Pronunciation variation in many ways is systematic, yielding patterns that a canny listener can exploit in order to aid perception. This work asks whether listeners actually do draw upon these patterns during speech perception. We focus in particular on a phenomenon known as paradigmatic enhancement, in which suffixes are phonetically enhanced in verbs which are frequent in their inflectional paradigms. In a set of four experiments, we found that listeners do not seem to attend to paradigmatic enhancement patterns. They do, however, attend to the distributional properties of a verb's inflectional paradigm when the experimental task encourages attention to sublexical detail, as is the case with phoneme monitoring (Experiment 1a-b). When tasks require more holistic lexical processing, as with lexical decision (Experiment 2), the effect of paradigmatic probability disappears. If stimuli are presented in full sentences, such that the surrounding context provides richer contextual and semantic information (Experiment 3), even otherwise robust influences like lexical frequency disappear. We propose that these findings are consistent with a perceptual system that is flexible, and devotes processing resources to exploiting only those patterns that provide a sufficient cognitive return on investment.

Keywords: Probability, perception, pronunciation variation, cognitive resources, phonetics, morphology

Spoken language is a labyrinth of variability, conveyed on an acoustic stream that often seems to carry more noise than signal. Yet that stream is rife with systematic patterns, and listeners are quick to exploit them to their advantage. In the realm of pure phonetics they use nasal coarticulation to predict upcoming nasal consonants (Beddor, McGowan, Boland, Coetzee, & Brasher, 2013); [ɹ]-coloring on preceding sonorants to predict upcoming rhotics (Heinrich, Flory, & Hawkins, 2010); stem duration to predict upcoming suffixes (Blazej & Cohen-Goldberg, 2015; Kemps, Ernestus, Schreuder, & Baayen, 2005; Kemps, Wurm, Ernestus, Schreuder, & Baayen, 2005); and syllable duration to predict upcoming word boundaries (Davis, Marslen-Wilson, & Gaskell, 2002; Salverda, Dahan, & McQueen, 2003). In the more abstract realm of distributional statistics, they use frequency distributions within the lexicon, within morphological families, and within inflectional paradigms to help identify and name words (Baayen, Levelt, Schreuder, & Ernestus, 2008; Baayen, Wurm, & Aycock, 2007; Moscoso Del Prado Martín, Kostić, & Baayen, 2004; Tabak, Schreuder, & Baayen, 2005, 2010). Listeners are canny and opportunistic, and capable of drawing upon many different types of patterns in the speech stream to aid perception and comprehension.

In this study, we explore more deeply the types of detail people use during perception, and further examine the limits of their abilities to use it. Specifically, we focus on the role of pronunciation variation in the identification of English verbs inflected with the third-person singular present tense suffix *-s*. This suffix is of particular interest because it reflects two types of relationships that listeners navigate when they are presented with a complete sentence. The first is the contextual relationship between the inflected verb and the syntactic subject that governs agreement—a relation that is enormously influential in

language processing. In eye-tracking experiments, for example, both adults and children use agreeing determiners or verbs to predict the identity of an upcoming noun (Dahan, Swingley, Tanenhaus, & Magnuson, 2000; Lukyanenko & Fisher, 2016), and ERP studies have observed a robust EEG response to sentences that violate expected number agreement relations (e.g., Molinaro, Barber, & Carreiras, 2011; Osterhout & Mobley, 1995).

The second relationship in which inflected verbs participate is paradigmatic, consisting of the set of morphological relations that hold between the word form itself and related forms of that same lexeme. As with contextual ties in sentence processing, paradigmatic ties also affect lexical processing. In picture-naming and lexical decision tasks, for example, both accuracy and reaction time are sensitive to inflectional entropy, an information-theoretic measure that reflects both the size of an inflectional paradigm and the frequency distribution of its members (Baayen et al., 2008; Tabak et al., 2005, 2010). In fact, not only does the entropy of a target word's inflectional paradigm affect processing, but so does the extent to which that entropy diverges from the mean inflectional entropy of all words in that lexical category (Baayen et al., 2008). These findings are clear evidence that lexical storage of inflected verbs includes a complex set of interconnections between the target words and other paradigmatically related forms. Our focus on inflected verbs thereore provides the opportunity to explore two domains—paradigmatic and contextual relations—in which listeners' attention to acoustic and distributional patterns can in principle aid perception in multiple ways.

Of course, in order for listeners to draw upon acoustic cues to facilitate perception, it is necessary for those cues to be present. And, indeed, systematic patterns of pronunciation variation do provide both contextual and morphological information.

Consider first contextual probability, which is overwhelmingly linked with phonetic reduction. Words which are more probable in the context of surrounding words tend to be shorter, with more reduced vowels, more flapped coronal stops, and less frequent vowel epenthesis (Bell et al., 2003; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Gregory, Raymond, Bell, Fosler-Lussier, & Jurafsky, 1999; Jurafsky, Bell, Gregory, & Raymond, 2001; Tily & Kuperman, 2012). Subparts of words also participate in this pattern, such that syllables which are more frequent in the context of surrounding syllables are shortened and have centralized vowels (Aylett & Turk, 2004, 2006). Further, contextual probability can be determined with respect to syntactic structures rather than immediately surrounding lexical items, leading to patterns in which words in and immediately preceding syntactically probable constructions are also subject to shortening and deletion (Gahl & Garnsey, 2004; Kuperman & Bresnan, 2012; Tily et al., 2009). Both of these patterns come together in recent work on pronunciation variation in agreement suffixes, which found that suffixes themselves show systematic pronunciation variation that reflects the probability of using that particular agreeing form in the context of the sentence's subject (Cohen, 2014, 2015). Taken together, these findings indicate that linguistic units—be they words, syllables, or, in the current study, suffixes—are phonetically reduced when they are likely to be used in a particular context.

Acoustic cues to morphological structure, by contrast, seem to follow a different pattern. Of most immediate interest here is a somewhat counterintuitive effect called *paradigmatic enhancement*: Where contextually probable forms show phonetic reduction, forms which are frequent within their morphological paradigms often show some type of phonetic enhancement of their affixes. In this pattern, interfixes of Dutch compounds are

lengthened (Kuperman, Pluymaekers, Ernestus, & Baayen, 2007); past tense suffixes of Dutch verbs are less likely to be deleted (Schuppler, Van Dommelen, Koreman, & Ernestus, 2012); third-person singular present-tense suffixes in English verbs are lengthened (Cohen, 2014); and neuter singular and plural past tense suffix in Russian verbs show more peripheralized vowels (Cohen, 2015).

The work presented here explores how listeners draw on these patterns of pronunciation variation—contextual reduction and paradigmatic enhancement—when they process inflected verbs. Experiments 1a, 1b, and 2 present target verbs in isolation in phoneme monitoring (Exp. 1a, 1b) and lexical decision tasks (Exp. 2). They focus on the role of paradigmatic enhancement. In particular these experiments explore the link between perception and pronunciation patterns in production. If words which are more frequent in their morphological paradigms are usually produced with lengthened affixes, then listeners will have become accustomed to hearing this pattern throughout their lives. Deviations from expected pronunciation patterns slow perception (Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005), so if listeners do exploit paradigmatic enhancement to aid perception, we would expect them to be faster to recognize paradigmatically probable word forms that are produced with the expected affixal lengthening, and faster to recognize the improbable words produced without lengthening. Yet on the other hand, forms which are paradigmatically probable are also those which are most likely to be selected from an inflectional paradigm, and this usage probability might well afford them a certain degree of robustness against phonetic variation. This expectation is supported by the findings of Ernestus & Baayen (2007, Exp. 1) who observed that the frequency of a word's stem interacted with phonetic reduction of the prefix in speech perception. Forms with high-

frequency stems showed little effect of prefix reduction on lexical decision reaction times, while those with low-frequency stems were much more sensitive. In the current study, if paradigmatically probable words are similarly insulated from the effect of affixal enhancement, then we would expect to see faster reaction times for higher-probability verb forms regardless of the phonetic realization of the suffixes.

Experiment 3 brings in an exploration of the role of contextual probability, by presenting target words in two different sentence frames. In one frame, the "use context," the words are used as main verbs in the sentence (e.g., *My grandfather bakes excellent pies*). Here, the subject-verb agreement relation renders the existence of the *-s* suffix on the verb entirely predictable and probable. In the other frame, the "mention context," words are mentioned in quotative or metalinguistic contexts (e.g., *He learned the word bakes in English class*), which deprive them of any contextual support that might lead listeners to expect an agreement suffix. If listeners draw on their knowledge of pronunciation patterns by which contextually probable linguistic units tend to be phonetically reduced, then we might expect them to respond more quickly to shortened suffixes than to lengthened suffixes in the use sentences, and vice versa in the mention sentences. Alternatively, if the more probable forms are insulated against phonetic variation, then listeners might respond more quickly to the use context sentences overall, regardless of suffix length.

Experiment 1a

Experiment 1a was designed to provide the listener with the greatest chance to detect departures from the paradigmatic enhancement effect. Participants completed a phoneme-monitoring task, in which the target phoneme was the sound [s], and critical stimuli were all third-person singular English verbs carrying the suffix *-s* (e.g., *looks, aches*). The

phoneme-monitoring task therefore ensured that listeners were highly alert to sublexical structure of the stimuli, while the use of the segment [s] as the target phoneme ensured that listeners were attentive to the phonetic properties of the suffixes on the target stimuli.

Methods

Participants.

Thirty-six UC Berkeley students and members of the surrounding UC Berkeley community were recruited to participate in this study (6 male). They ranged in age from 18 to 41 (mean 19.7), all spoke English as their first language, and none reported any hearing problems.

Materials.

The critical stimuli comprised 50 one-syllable English verbs. All verbs had stems ending in either [p] or [k] to ensure that the final suffix was a voiceless [s] that could be easily segmented from the release burst of the preceding stop. No verb contained the phoneme [s] anywhere except in the suffix. Frequency values for all members of the critical verbs' inflectional paradigms were extracted from the part-of-speech tagged SUBTLEX-US corpus (Brysbaert, New, & Keuleers, 2012). Lexical frequency was represented by the log-transformed frequency of the third-person singular form, and paradigmatic probability was calculated by dividing each verb's third person singular frequency by the summed frequencies of all possible verb forms in the lexeme. Because we were concerned specifically with the role of the verbal inflectional paradigm on processing, we extracted only the verb-specific frequency measures from SUBTLEX. Homophonous usages of the word forms as nouns or other parts of speech were not included in the frequency measures. The verbs ranged in paradigmatic probability from a minimum of 0.005 (*copes*) to 0.496 (*reeks*). These values were log-transformed to reduce skew, and then centered by

subtracting from each value the mean of all log-transformed relative frequencies.

The verbs were recorded in one sitting three times each by a female native speaker of English, and the best token, with the smallest degree of creak or other phonetic artifacts, was selected from each set of three repetitions. Each verb's raw suffix duration and raw stem duration was extracted, and the ratio of these durations calculated. The average suffix-to-stem duration ratio over all raw recordings was about 0.64. In order remove any paradigmatic enhancement pattern in the raw recordings that the speaker might have unconsciously produced, these words were then adjusted by Praat script (Boersma & Weenink, 2015) so that for every word the ratio of suffix duration to stem duration was 0.64, matching the average ratio across the whole set. These recordings make up the "norm" condition. Next, for each verb a second version was constructed, such that the suffix duration was reduced by 25% from the norm condition, producing the "short" condition. A third version was also produced, with the suffix lengthened by 25% of the normalized duration, producing the "long" condition. Finally, all recordings were adjusted in amplitude by RMS-based normalization using the Ffmpeg software (http://www.ffmpeg.org), such that the mean amplitude was equal across all the audio files.

In addition to the 150 critical stimuli—three audio versions for each of the 50 critical verbs—a further 250 non-critical stimuli were recorded in the same session. These stimuli were not manipulated in any way beyond normalizing the amplitude to match the critical stimuli.

Design.

Three experimental lists were constructed and rotated across subjects. In each list the critical verb appeared in two length conditions, for a total of 100 critical stimuli in each list.

For example, in list A, the verb *aches* appeared in the short and norm condition, list B contained *aches* in norm and long condition, and list C contained *aches* in the short and long condition. These length-pairings were rotated across all fifty stimuli and over the three lists, so that on each list the three possible length-pairings appeared either 16 or 17 times, and each verb appeared in a different length-pairing across the three lists.

In addition to the 100 critical stimuli, each list contained 300 fillers, 100 of which shared a stem with the critical verbs, and 200 of which contained a different stem. The 100 fillers sharing a stem with the critical items were different inflectional forms of the verb, while the latter 200 represented three inflectional or derivational forms of 50 filler stems. The resulting design ensured that, regardless of whether a stem appeared as a critical stimulus (e.g., the stem *ache*) or only as a filler (e.g., the stem *build*), participants heard each stem four times, with a parallel distribution of forms: Two tokens were repetitions (e.g., the two length conditions of the critical stimulus *aches* or a straightforward repetition of the filler *builder*), and two were different inflectional forms of that stem (e.g., *ached*, *aching* and *build*, *built*). Out of the 400 stimuli in each list, slightly less than half (174) contained an [s], of which 100 were the 50 critical stimuli in the two length conditions, and the remaining 74 were fillers.

Procedure.

The experiment was built using the OpenSesame experimental software (Mathôt, Schreij, & Theeuwes, 2012) on a desktop computer running the Linux operating system. Stimuli were presented binaurally through headphones set at a comfortable volume, in a different random order for each participant. Responses were recorded on a serial response button box, with the left-most button used to indicate that [s] was present, and the right-most button used to

indicate that [s] was not present. After filling out a brief language background questionnaire, participants were shown into the booth. The experiment started with 10 practice trials, and participants had a chance to rest or ask questions after the practice block, and every 50 trials of the full experiment. The experiment lasted about 30 minutes, and participants were compensated $5.00.

Results

Data from one of the participants was discarded because an experimental error prevented the stimuli from being displayed in random order. Of the 3500 observations provided by the remaining 35 participants, 12 observations timed out before a response could be made, and 70 had a reaction time of more than 1700 ms, and were discarded for excessive slowness. Of the remaining 3418 observations, 201 were incorrect, for an accuracy rate close to ceiling, at 94.1%.

Accuracy.

Response accuracy was analyzed with mixed-effects logistic regression modeling using the lme4 package (version 1.1_7; Bates, Maechler, Bolker, & Walker, 2015) in the R programming environment (version 3.1.3, R Core Team, 2015). Since the high accuracy rate and smaller sample size limited the complexity of the model, the random effects structure could accommodate only intercepts for subject and word, and the only fixed effect besides length and paradigmatic probability that could be incorporated was lexical frequency. The contribution of each fixed-effect predictor to model fit was assessed through a log-likelihood ratio test comparing the model containing each predictor of interest to a simplified version of the model lacking only that predictor. Accurate responses were more likely with high-frequency verbs ($\beta = 0.090$, $SE(\beta) = 0.037$, $\chi^2(1) = 5.56$, $p < 0.05$), but

adding length, paradigmatic probability, and their interaction did not improve model fit over the simpler model containing only frequency ($\chi^2(5) = 3.72$, $p > 0.1$), and neither did adding either length ($\chi^2(2) = 3.50$, $p > 0.1$) or paradigmatic probability ($\chi^2(1) = 0.03$, $p > 0.1$) as simple effects by themselves.

Reaction time.

Reaction times (RTs) of the 3218 correct responses were measured from the onset of the audio stimulus, log-transformed to reduce skew, and submitted to a mixed effects regression analysis. Control predictors were added first in order to account for as much variability in response time as possible. The critical predictors of length, paradigmatic probability, and their interaction, were added last. Control predictors included the duration of the stem up to the onset of the [s], the log-transformed reaction time on the previous trial, the trial number, and the log-transformed frequency of the target word. These control predictors were examined for correlations with the critical predictor of paradigmatic probability, and Table 1 shows that the correlations were acceptably low.

Insert Table 1 about here

The amount of data was insufficient to allow a maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013), and indeed recent approaches to mixed effects models have argued that maximal models can overfit the data at the loss of statistical power (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). We determined which random slopes to include by evaluating whether individual slopes improved model fit. According to a log-likelihood ratio test, random slopes were not justified for the analysis of reaction time, and so the only random effects in the final model were intercepts for subject and word.

Table 2b summarizes how model fit was improved by the addition of the critical predictors of length and paradigmatic probability. Leaving out any interactions, the inclusion of length in a model that already contained paradigmatic probability significantly improved model fit ($\chi^2(2) = 33.59, p < 0.001$), while the inclusion of paradigmatic probability in a model that already included length did not ($\chi^2(1) = 0.283, p = 0.595$). The interaction of paradigmatic probability and length, however, did improve model fit over the simpler model lacking paradigmatic probability in any form ($\chi^2(3) = 8.644, p < 0.05$). The nature of this effect, illustrated in Figure 1, was that the slope of paradigmatic probability for norm duration suffixes was not significantly greater than 0 ($\beta = 0.0083$, $SE(\beta) = 0.0056$, $t = 1.47$), while shortened and lengthened suffixes had significantly more negative slopes, consistent with faster RTs for forms with higher paradigmatic probabilities (short $\beta = -0.0149$, $SE(\beta) = 0.0064$, $t = -2.32$; long $\beta = -0.0170$, $SE(\beta) = 0,.0064, ,$ $t = -2.65$). A full summary of the final model is given in Table 2a.

Insert Table 2a-b about here

Discussion

Experiment 1a was designed to determine whether listeners are sensitive to the observed pronunciation pattern whereby paradigmatically probable verbs have lengthened suffixes. If they are sensitive to it, and draw on it to aid processing of acoustic stimuli, then an interaction should have resulted: listeners would respond more quickly to words with high paradigmatic probability of the suffixes were long, and more slowly if suffixes were short. If they do not draw on this pronunciation pattern to aid perception, then higher paradigmatic probability was predicted to aid processing regardless of the phonetic realization of the suffix.

The results seem to show that listeners are indeed sensitive to some relation between suffix duration and paradigmatic probability, as indicated by the interaction between those two factors in the regression model. The nature of their sensitivity, however, does not straightforwardly support either of the two predictions. Compared to its effect on norm-length suffixes, increased paradigmatic probability has a facilitative effect on long suffixes, consistent with the first prediction. Yet it had the same effect in short suffixes, contrary to that prediction.

Because these results are difficult to explain, we replicated Experiment 1a using a larger set of verbs and a larger subject pool, to see whether the unpredicted interaction would persist.

Experiment 1b

Experiment 1b was designed to replicate Experiment 1a, differing only in that both the stimulus set and the subject pool were increased by 50%.

Methods

Participants.

Fifty-seven students and members of the George Washington University campus community participated in Experiment 1b (15 male). Participants were compensated with either course credit or with $10 depending on the recruitment source. The subjects ranged from 18 to 47 in age (median 22), spoke English as their first language, and reported no hearing disorder.

Materials.

Materials were again one-syllable English verbs with stems ending in /p/ and /k/, and not containing the phoneme /s/ anywhere except in the third-person singular suffix. Seventy-

five verbs were used this time, and paradigmatic probability was calculated as in Experiment 1a. The minimum relative frequency of the form within the inflectional paradigm ranged from .004 (*hikes*) to .723 (*creaks*). As before, these values were log-transformed and mean-centered. All critical words and fillers were recorded by the same speaker as in Experiment 1a. The suffix length-manipulation were identical to the procedure in Experiment 1a. The only difference in preparing the stimuli consisted of normalizing the intensity of all critical stems and fillers in Praat, rather than through a separate program.

Design.

The design was identical to Experiment 1a, differing only in the larger number stimuli. To balance out the 50% increase in critical stimuli, we also increased the number of fillers by the same amount, and maintained the balance of repetition across forms and stems between critical and filler stimuli. Final experimental lists contained 600 stimuli in all, of which 150 critical trials.

Procedure.

The experimental procedure was carried out using E-Prime (ver. 2.0) running on a Windows computer, with responses recorded on a Chronos device and a headphone (Sony MDRV700) attached to the Chronos response box. The overall procedure was identical to Experiment 1a, except that participants were given verbal instruction without any practice block in the beginning of the experiment. Participants were given the opportunity to rest or ask questions every 100 trials of the full experiment. The total duration of the study per participant was approximately 45 minutes.

Results

The 57 participants provided 8550 critical observations. Based on visual inspection of the density curve of all reaction times, RTs slower than 450 ms or faster than 1600 were discarded as outliers. Trials with no response, or trials following trials with no response, were also discarded. This resulted in the loss of 149 observations (1.75% of data) for the accuracy analysis, leaving 8401 responses for the accuracy analysis. The correct 7902 responses were then retained for the RT analysis.

Accuracy.

As in Experiment 1a, accuracy was modeled with mixed effects logistic regression, with random intercepts for subject and word. Also as in Experiment 1a, responses were more accurate with higher frequency verbs (β = 0.094, $SE(\beta)$ = 0.032, z = 2.93, $p < .01$), but frequency was the only predictor that affected accuracy ($\chi^2(1) = 8.06$, $p < 0.01$). Adding length, paradigmatic probability, and their interaction did not improve model fit over the simpler model containing only frequency ($\chi^2(5) = 3.80$, p =.58), and neither did adding either length ($\chi^2(2) = 1.13$, p = .57) or paradigmatic probability ($\chi^2(1) = 0.09$, p =.77) as simple effects by themselves.

Reaction time.

Because Experiment 1b was designed to be a replication of Experiment 1a, with a specific focus of examining the role of the interaction between length and paradigmatic probability, we did not build up the full model through forward addition of predictors, but rather started with the final fixed effects structure from the model in Experiment 1a. As in Experiment 1a, we confirmed that the numerical predictors were not highly correlated with paradigmatic probability (Table 3). Unlike Experiment 1a, the larger data set allowed the inclusion of random slopes for the interaction of length and paradigmatic probability by subject, and

length by verb. We report the model with the more maximal random effects structure here, although the pattern of results was similar when we looked only at models containing random intercepts.

Insert Table 3 here

In this experiment, the interaction between length and paradigmatic probability was not significant, and removing the interaction did not significantly affect the goodness of fit of the model ($\chi^2(2) = 3.006$, $p = .22$). In the absence of any interaction, however, higher paradigmatic probability still had a facilitative effect across all three suffix lengths ($\beta = -0.006$, $SE(\beta) = .002$, $t = -2.64$). A full model summary is given in Table 4.

Insert Table 4 about here

Discussion

Experiment 1b replicated Experiment 1a with more subjects and items. Although most effects remained quite similar, in this replication there was no interaction between stem duration and paradigmatic probability. Rather than the facilitative effect appearing to apply only to shortened and lengthened stem durations, as in Experiment 1a, here it applied to all stem durations similarly, suggesting that the emergence of a significant interaction in the smaller Experiment 1a did not reflect a real effect.

The simple effects of lexical frequency and paradigmatic probability did, however, show that listeners engaged in at least some degree of lexical processing, and were not simply mindlessly monitoring for the phoneme /s/. Although it is intuitively straightforward that words with high lexical frequency are recognized quickly, how is it that recognition of words with higher paradigmatic probability was also speeded? One possible explanation for the facilitative effect of paradigmatic probability is based on previous findings showing that

listeners perceive speech incrementally, narrowing down the set of possible forms for a target word as the acoustic stream unfolds (Allopenna, Magnuson, & Tanenhaus, 1998; Magnuson, Dixon, Tanenhaus, & Aslin, 2007; Van Petten, Coulson, Rubin, Plante, & Parks, 1999). In this way, the acoustic information in the stem allowed listeners in this experiment to identify the likely morphological paradigm of the word in question, and access the distributional information that characterizes it, before the onset of the suffix. If the third-singular, *s*-suffixed form is more frequent in that paradigm, then listeners could have been prepared for the [s], and thus readier to respond to the acoustic cues when they arrived. Under this interpretation, the lack of an interaction between suffix length and paradigmatic probability simply reflects the fact that listeners do not attend to the match between an acoustic stimulus and probability-conditioned variation in the suffix duration. By the time they have heard even part of the suffix, they have enough information to identify the word and recognize that it contains the target phoneme without needing to attend to lower-level subphonemic cues. Indeed, the faster RT for the short suffixes compared to long suffixes both here and in Experiment 1a support this view: Listeners do not need more acoustic information to identify the [s], and seem instead to respond faster as a function of how quickly the stimulus finishes playing.

Alternatively, it could be that listeners actually are capable of hearing mismatches, but this particular experiment was insufficiently sensitive to detect it. Kemps and colleagues (Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005) showed that as far as *stem* duration is concerned, listeners are quite good at recognizing when the expected durational pattern does not match what they expect, and respond more slowly on number-decision and lexical decision tasks in both Dutch and English if the stem duration is

lengthened or shortened beyond expected prosodic patterns. However, there are several differences between those studies and the current work. For example, Kemps and colleagues manipulated the duration of the stem, while we focused on the suffix; and the durational pattern Kemps and colleagues manipulated was not conditioned by paradigmatic probability, but by much more robust polysyllabic shortening (Lehiste, 1972). But, also, the number decision tasks (Kemps, Ernestus, et al., 2005) and lexical decision tasks (Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005) would have required a greater degree of lexical processing than the low-level phoneme monitoring used here. To be sure, the frequency and paradigmatic probability effects in this work show that listeners were engaging in some degree of lexical processing, but perhaps that level of engagement with the stimuli could be increased. Perhaps a task with a greater degree of lexical processing will allow listeners to more fully retrieve the pronunciation patterns associated with the distributional patterns in a morphological paradigm. If so, then in an experiment featuring a lexical decision task, we should see an interaction between suffix length and paradigmatic probability, such that RT is faster when paradigmatically probable forms have lengthened suffixes, and improbable forms have shortened suffixes. If, however, listeners' insensitivity to the relationship between suffix length and paradigmatic probability is unrelated to the degree of lexical processing, then we predict only another simple facilitative effect of paradigmatic probability.

Experiment 2

In order to determine whether a task that supports lexical-level processing induced an interaction between paradigmatic probability and suffix length, Experiment 2 used similar critical stimuli to Experiments 1a-b, but featured a lexical decision task, rather than a

phoneme-monitoring task.

Methods

Participants.

Forty-one UC Berkeley students and members of the surrounding UC Berkeley community were recruited to participate in this study (11 male). Two reported auditory problems and their results were therefore excluded from analysis. The remaining 39 participants (9 male) ranged in age from 18 to 62 (mean 22.0), and all reported English as their native language. None had participated in Experiments 1a-b.

Materials.

Since Experiment 2 was a lexical decision task, the same fifty critical verbs from Experiment 1a were used, along with 300 fillers. Because the fillers needed to include non-words for the lexical decision task, all stimuli, fillers and critical alike, were recorded anew in a single session by a female native speaker of English. As before, the norm condition of the critical stimuli was created by adjusting durations of the suffixes to match the mean proportion of the stem duration observed in the set of critical stimuli as a whole; the short condition was created by shortening those normalized suffixes by 25%; and the long condition was created by lengthening them by 25%. Filler words were not manipulated for length, but all stimuli, critical and filler alike, were amplitude-adjusted as in Experiment 1a, to ensure the same mean amplitude across all recordings.

Design.

As in Experiments 1a-b, the stimulus lists were constructed by presenting each verb in two of the three possible length conditions, with the three possible pairings rotated across the verbs and counterbalanced over three lists. This created 100 critical stimuli per list. Also as

in Experiments 1a-b, the fillers were constructed to mitigate the repetition of the critical stimuli, as follows. Of the 300 filler tokens, 100 were made up of the same 50 stems used in the critical stimuli, presented in two different forms. One form was a real word (e.g, *broke* or *chirped* to go with *breaks* and *chirps*) and one was a fake word. Some of the fake words modified the stem directly (e.g. *chirb*), and some contained a false continuation (e.g., *brokem*). This ensured that participants would need to listen carefully to what followed the stem in order to determine whether a given stimulus was a real word or not. These fillers, together with the 100 critical stimuli, meant that half of each 400-word list was built around fifty stems, with four tokens representing three types for each stem.

The remaining 200 filler tokens were constructed in parallel to the first 200, around an additional 50 filler stems. Again, each stem appeared in four tokens representing three types. Two of these appearances were identical repetitions, mirroring the repetition of the critical stimuli in the different length-pairings of the critical stimuli, while the other two appearances were different forms. At least one variant for each filler stem was a non-word, and sometimes a given filler stem (e.g., *plow*) contributed two non-word forms (e.g., a false-suffix form *plowish* and a stem-changed form *plar*).

This design ensured that, across all 400 stimuli, each of the 100 stems (50 critical, 50 filler) appeared a total of four times in three different forms, of which at least one and as many as three could be non-words. Overall, 133 forms (33%) were non-words. These stimuli were presented in a different random order for each participant.

Procedure.

The procedure was identical to Experiments 1a-b, differing only in that participants used their dominant hand to respond “yes” if the stimulus was a real word.

Results

Of the 3900 responses to critical stimuli provided by the 39 participants whose results were retained, 115 were excluded for excessively slow responses (greater than 1700 ms), time-outs, time-outs on the previous trial, responding before the end of the stimulus, and illegal responses (i.e., selecting a button that did not correspond to "yes" or "no" on the button-box). Of the remaining 3785 responses, 161 were incorrect, for an accuracy rate of 95.7%.

Accuracy.

Accuracy was analyzed with mixed-effects logistic regression, using the same R installation and package specifications as in Experiment 1a. Also as in Experiment 1a (and 1b), the only predictor that improved the model fit of response accuracy was frequency: Accurate responses were more likely with high-frequency verbs ($\beta = 0.358$, $SE(\beta) = 0.655$, $\chi^2(1) = 22.23$, $p < 0.001$). Model fit was not improved by the addition of length to a model containing frequency and paradigmatic probability ($\chi^2(2) = 1.10$, $p > 0.1$), nor by the addition of paradigmatic probability to the model containing frequency and length ($\chi^2(1) = 2.66$, $p > 0.1$), nor by the addition of both critical predictors and their interaction to the model containing only frequency ($\chi^2(5) = 6.21$, $p > 0.1$).

Reaction time.

Reaction times were measured from the onset of the audio stimulus. Visual inspection of the distribution of reaction times showed no appreciable skew, and so RT values were not log-transformed, but used as raw durations in milliseconds. The RTs of the 3624 correct responses were submitted to a mixed-effects linear regression analysis in the same manner as in Experiment 1a. The control predictors included the duration of the audio stimulus, the reaction time of the previous trial, the trial number, and the lexical frequency of the

particular stimulus. Table 5 shows that these variables were acceptably uncorrelated with paradigmatic probability. As in Experiment 1, random effects included intercepts for subject and word.

Insert Table 5 about here

As Table 6b summarizes, adding suffix length to the model containing only control predictors and paradigmatic probability significantly improved model fit ($\chi^2(2) = 17.83$, $p < 0.001$). Compared to stimuli with short suffixes, participants responded faster to stimuli with norm or long suffixes (norm $\beta = -24.65$, $SE(\beta) = 6.33$, $t = -3.85$; long $\beta = -30.70$, $SE(\beta) = 7.98$, $t = -3.85$). There was no improvement of model fit when paradigmatic probability was added as a predictor, either alone ($\chi^2(1) = 0.20$, $p > 0.1$) or in interaction with length ($\chi^2(3) = 0.30$, $p > 0.1$). This lack of interaction is illustrated in Figure 2. The parameters of this full model are summarized in Table 6a.

Insert Figure 2 about here

Insert Table 6a-b

Discussion

Experiment 2 employed a lexical decision task, rather than a phoneme-monitoring task, in order to encourage more lexical-level processing than Experiments 1a-b. Counter to both predictions proposed in the Discussion for Experiment 1b, not only did the facilitative effects of paradigmatic probability fail to interact with length, but in fact all effects of paradigmatic probability disappeared entirely. This is unlikely to be due to experimental error, as control predictors behaved as expected. The coefficient for file duration, for example, was extremely close to 1, which reflects the fact that every one-millisecond increase in file duration leads to approximately a one-millisecond increase in RT—an

entirely natural consequence of measuring RT from the onset of the stimulus. The positive effect of RT from the previous trial reflects the tendency for participants to respond at fairly consistent rates throughout the experiment, so a slower or faster previous trial would yield a slower or faster RT on the current trial (although see H. Baayen, Vasishth, Kliegl, & Bates, 2017, for a more nuanced view). The faster RTs for norm and long stimuli are also straightforward: Participants can make lexical decisions faster when they have more acoustic information available to identify the stimulus. And, crucially, the facilitative effect of lexical frequency, combined with the high accuracy rate on the lexical decision task, indicates that participants are indeed performing some degree of lexical processing.

One explanation for the disappearance of the effect of paradigmatic probability is rooted in the observation that the details of an experimental task can affect not only the level of processing that participants engage in, but also the types of information that they draw on to perform the task. Baayen et al. (2007), for example, observed that participants draw on morphological properties such as family size and inflectional entropy in visual lexical decision tasks, but not in visual naming tasks or, crucially, auditory lexical decision tasks. Since Experiment 2 was also an auditory lexical decision task, then perhaps the absence of any paradigmatic probability effect is reflecting the same reduction of morphological processing that Baayen et al. observed. Although we had used a lexical decision task in the hopes that a greater degree of lexical processing would lead to a stronger morphological awareness, it is also true that a lexical decision task requires participants to process the entirety of the word, while a phoneme-monitoring task requires them to be aware of sublexical constituents. In the case of suffixed verbs, the key sublexical constituent—the target phoneme [s]—was in fact a complete morpheme, and thus the

nature of the task in Experiments 1a-b could actually have heightened attention to morphological constituency. In Experiment 2, by contrast, participants need to abstract attention away from subparts of the words, and instead consider the identity of each stimulus as a whole. For the more complicated morphological properties, such as paradigmatic probability, this shift in attention away from the subcomponents of the auditory stimuli could have obscured any detectable effect of paradigmatic probability on perception. If this is so, then other types of auditory lexical decision tasks should equally well require participants to identify words as whole entities, and thus there should be no reappearance of an effect of paradigmatic probability on reaction times.

The account proposed above is based on the notion that attentional resources are limited, and when they are transferred from one aspect of a stimulus (such as morphological constituency) to another aspect (such as lexical status), sensitivity to these properties of the stimulus changes accordingly. A more general implication of this account is that in more complex tasks, processing resources that might otherwise be available for drawing on subtle lexical patterns to aid perception are required instead to complete the tasks. As a result, effects of only the most robust lexical properties—those that provide the greatest return for the investment of cognitive resources—can be observed. In a phoneme monitoring task in which the target phoneme is also a complete morpheme, information about morphological properties like paradigmatic probability and suffix duration might aid completion of the task, and thus justify the expenditure of resources required to process and incorporate that information into the perceptual task. In a lexical decision task, the expenditure is not justified, and so participants do not draw on that information. As a result, no effect of paradigmatic probability emerges.

If this account is accurate, then as more informative cues become available to participants, participants might cease to use even usually robust cues to word identity, such as lexical frequency, in favor of the more informative cues provided in the different task. Experiment 3 tests these predictions by asking participants to complete another lexical decision task in which target words are embedded in a full sentence.

Experiment 3

In Experiment 3, participants completed a lexical decision task on the same set of words that were used in Experiment 2, only in this case they were embedded within a complete sentence. Two types of sentences were considered. The first type was a "mention" context, i.e. the critical word appeared as a metalinguistic topic of discussion, as in (1) below. The second type was a "use" context, in which the critical word was used as an integrated component of the sentence—in this case, as the main verb, agreeing with a singular subject, as in (2) below.

(1) a. She recalled the word *dips* instantly.

b. He memorized the word *dips* in two seconds.

(2) a. My friend always *dips* crackers in sour cream.

b. The child never *dips* carrots in hummus.

The mention sentence provided no contextual clues as to the inflectional form of the target word, thus ensuring that the probability of observing a singular agreement suffix was minimized. By contrast, the use sentence provided a singular subject as the governor of verbal agreement, thus ensuring that a singular agreement suffix on the main verb was obligatory.

Experiment 3 therefore makes it possible to answer three questions. The first,

discussed in the introduction to this paper, is whether listeners make use of production patterns to aid perception. If they do, then they should respond more quickly to shortened suffixes when they are contextually probable (i.e., in the use sentences), and more quickly to the lengthened suffixes when they are contextually improbable (i.e., in the mention sentences). If, instead, listeners simply respond more quickly to any probable word form, regardless of how it matches or does not match observed phonetic reduction patterns, then listeners should respond more quickly in the contexts in which the third singular verb form is probable, the use sentences, compared to the contexts in which that form is unpredictable, the mention sentences.

The remaining two questions come from the Discussion of Experiment 2. If the absence of an effect of paradigmatic probability in Experiment 2 was due to the whole-word processing encouraged by a lexical decision task, then the absence should persist in the current experiment. Further, if the additional cues to word identity and lexicality provided by the surrounding sentence context are so useful that other lexical properties like word frequency are insufficiently informative to justify the cognitive resources required to incorporate them into task responses, then the effect of frequency should weaken or disappear, especially in the use context sentences.

Methods

Participants.

Thirty-nine participants were recruited from the same community as Experiments 1a and 2. Data from one participant was excluded from analysis due to self-reported severe hearing loss in one ear. The remaining 38 participants (24 F, mean age 22.6), all reported English as either their first language, or else as their dominant language in case they had started

speaking as heritage speakers of another language. None reported learning English later than age 6, and the mean age of acquisition of English was 1 year old. No participants had participated in Experiments 1a-b or 2.

Materials.

Forty verbs were selected from the set used for Experiments 1 and 2, and four sentences were written for each verb. Two sentences placed the verb in a mention context, such that the word appeared as a metalinguistic topic of discussion in the sentence as in (1a-b) above, while the other two sentences placed the word in a use context, so that the word was incorporated naturalistically into the structure of the sentence, as in (2a-b) above. In each sentence, verb suffix duration was manipulated in the same way as in Experiments 1 and 2 above, creating the same long, norm, and short length conditions.

Each critical verb was also paired with a non-word form, as in Experiment 2, which was constructed either by adding a fake suffix to the end of the stem, or by changing the stem by one phoneme. This non-word variant of the critical verb always appeared in a use context, as in (3) below.

(3) The boy and his *dippish* rode their bikes together.

In addition to the 40 critical verb stems, an additional 40 stems were selected to appear as fillers. Each filler stem had either two or three different forms, of which one was real (e.g., *steamed*, *shirt*) and the other(s) fake (e.g., *steamish*, *shirp*). Three sentences were then constructed for each filler stem. For half of these filler stems, all of the sentences were mention context. For the other half, two of the three sentences were use context, and the third was mention context. These sentences therefore made up a set of 120 filler sentences, of which 80 were mention context, and 40 were use context.

Design.

Six experimental lists were designed, and the three different lengths and two different sentence contexts were rotated across the lists. Each experimental list contained 120 sentences built around a critical stem, of which 80 contained the target suffixed verb, half appearing in use context sentences and half in mention context, and the remaining 40 all contained non-words in a use context. The fillers balanced out the distribution of sentence contexts, such that of the 120 filler sentences, 80 were mention context sentences, and only 40 were use context sentences. Eighty of the filler items were non-words, while 40 were real words. Thus each list contained 240 sentences, half of them use context and the other half mention context, and half of the words were real words, while the other half were non-words.

Within a list, the target verb appeared twice, with two different suffix lengths, as in Experiments 1 and 2. Thus, in lists A1 and A2, *aches* might appear in the long and norm conditions; in lists B1 and B2 it appeared in norm and short; and in lists C1 and C2 in appeared in short and long. Further, in a given list each verb appeared either in only use-context sentences or only mention-context sentences. Thus, in lists A1, B1, and C1 the verb *aches* appeared in its two use context sentences, while in lists A2, B2, and C2 it appeared in its two mention context sentences.

This design ensured an equal number of use context and mention context sentences within each experimental list, while minimizing the informativeness of any predictive strategies. Any given stem could appear in both contexts, or only one, and the three appearances of each stem could include repetitions, or each appearance could be a completely new form. In this way the different conditions of the critical verbs were evenly

distributed across the lists, but the distribution of the fillers provided sufficient unpredictability to prevent participants from using any sort of process-of-elimination strategy to predict the identity or lexicality of a given stimulus.

Procedure.

Experiment 3 was carried out in much the same way as Experiments 2. The only difference was that for each trial participants also saw a visual display of the sentences, with a blank spot in the position of the target word, as in (4) below:

(4) She recalled the word _____ instantly.

Participants were instructed to press the "yes" button if the word that was used in the audio recording in the position of the blank spot was a real word, and the "no" button otherwise. Participants were also instructed to respond as quickly as possible, without waiting for the end of the sentence. The full experiment took about 45 minutes, and participants were compensated $7.50.

Results

Of the 3040 responses collected from the 38 participants, 164 were removed for excessively fast or slow response time (greater than 1700 ms or less than 200 ms); invalid responses (i.e., pressing a button other than the two allowed responses on the button box); and time-outs on the previous trial. Of the remaining 2876 responses, 150 were incorrect, for an accuracy rate of 94.8%.

Accuracy.

As in Experiments 1 and 2, accuracy was analyzed with a mixed effects logistic regression model. Random effects were restricted to intercepts for subject and sentence, with a random slope for lexical frequency by subject. Interactions between length and context, and length

and paradigmatic probability, all failed to improve model fit against the simpler model lacking the interactions (length:context χ^2 (2) = 0.54, $p > 0.1$; length:paradigmatic probability χ^2 (2) = 2.00, $p > 0.1$). Further analysis, summarized in Table 7b, accordingly considered only the result of adding critical predictors as simple effects to models that lacked them entirely.

The addition of both length and context as simple predictors improved model fit compared to the respective simpler models lacking them (length χ^2 (2) = 6.15, $p < 0.05$; context χ^2 (1) = 66.9, $p < 0.001$). The addition of log-transformed lexical frequency trended towards improved model fit (χ^2 (1) = 3.55, p =0.06), and was retained in the final model. When paradigmatic probability was added, however, model fit did not improve (χ^2 (1) = 1.13, $p > 0.1$).

Coefficient estimates for length and context indicate that verbs with normalized suffixes elicited more accurate responses than shortened or lengthened suffixes (β = 0.45, *SE(β)* = 0.22, z = 2.02, $p < 0.05$), and verbs in use contexts elicited more accurate responses than verbs in mention contexts (β = 1.58, *SE(β)* = 0.21, z = 7.55, $p < 0.001$). According to the marginal effect of lexical frequency, accuracy increased with higher lexical frequency (β = 0.10, *SE(β)* = 0.05, z = 1.97, $p < 0.05$). The full model summary is reported in Table 7a.

Insert Table 7a-b about here

Reaction time.

As in Experiment 2, the reaction times (RTs) were measured from the onset of the target word, and visual inspection of the distribution of RTs revealed no skew. As a result, RTs were not log-transformed, but analyzed in their raw form, in ms. The 2726 correct

responses were submitted to a mixed effects regression analysis modeling reaction time as the dependent variable.

Control predictors in the model included the trial number, the reaction time on the previous trial, and the raw duration of the target word as measured in milliseconds. Critical predictors tested in this model included sentence context, suffix length, log-transformed lexical frequency, and paradigmatic probability. Table 8 shows that the numerical control predictors were not highly correlated with lexical frequency and paradigmatic probability. Random effects included intercepts for participant and sentence frame, with a random slope for context by participanbt. Additional random slopes for the other critical predictors did not improve model fit.

Insert Table 8 here

As with the accuracy analysis, the interactions of length with paradigmatic probability and context failed to improve model fit. Retaining them would have produced an overfitted model, obscuring possible effects of other predictors (length:paradigmatic probability $\chi^2(2) = 3.01$, $p > 0.1$; length:context $\chi^2(2) = 2.84$, $p > 0.1$). Further analysis excluded therefore those interactions, and examined only the simple effects of the critical predictors.

Of the critical predictors of interest—paradigmatic probability, suffix length, sentence context, and lexical frequency—only sentence context improved model fit (See Table 9b). Compared to mention contexts, targets in use context sentences also elicited faster RTs ($\beta = -125.46$, $SE(\beta) = 10.90$, $t = -11.51$; $\chi^2(1) = 69.45$, $p < 0.001$). The inclusion of paradigmatic probability did not improve model fit ($\chi^2(1) = 0.08$, $p > 0.1$), and neither did suffix length ($\chi^2(2) = 0.69$, $p > 0.1$) or lexical frequency ($\chi^2(1) = 0.85$, $p > 0.1$). The full

parameter summary for this model is given in Table 9a.

Insert Table 9a-b about here

Discussion

The results of Experiment 3 fell into three broad patterns. First, there was no effect of paradigmatic probability, parallel to the absence observed in Experiment 2. The account proposed in the Discussion to Experiment 2 was that the nature of auditory lexical decision tasks requires listeners to process stimuli holistically, without attending to their sublexical structure. As a result, any lexical decision task will not show as strong effects of morphological distribution properties like paradigmatic probability, since words are less likely to be processed with respect to their components parts. The continued absence of any effect of paradigmatic probability in Experiment 3 is consistent with this account: Although the stimuli were full sentences and the procedure to identify the targets more complicated, the task was still a lexical decision task at heart, and therefore required participants to process words holistically, rather than attending to sublexical properties.

The absence of a paradigmatic probability effect is also consistent with the broader implications of a return-on-investment account. This account proposes that listeners allot cognitive resources to process linguistic information with an eye towards which information is most helpful in decoding the speech signal. In the presence of more useful linguistic cues, properties that affect recognition of words in some conditions might not justify the cognitive resource investment required to process them in other contexts. The syntactic and semantic information contained in a full sentence is so rich that it is not surprising to find listeners shifting their cognitive resources away from morphological processing. In our study that attentional shift meant that listeners ceased attending to

paradigmatic probability, but other types of morphological processing have also been shown to be affected. Hyönä, Vainio, & Laine (2002), for example, found that effects of morphological complexity in visual word recognition disappeared when target words were presented in sentence contexts.

The second result pattern in Experiment 3 was the weakening or absence of any effect of lexical frequency. Although the parameter estimate for frequency in the accuracy model did diverge significantly from 0, its inclusion in the model failed to significantly improve model fit, and the reaction time data showed absolutely no evidence of even a marginal effect of lexical frequency. This absence of a frequency effect could simply be due to the fact that participants had sufficient time to consider the sentence context and predict the upcoming word that would fill the gap before they heard it.[1] However, even if that is what they were doing, such a strategy is also consistent with the cognitive return on investment account. Listeners interpret speech incrementally, and make predictions about what they are about to hear (Altmann & Kamide, 1999). The more useful information they have to make their predictions about what is most probable, the less they need to draw on prior probabilities, such as lexical frequency.

This trade-off can also explain the disappearance of a suffix-length effect in Experiment 3. In Experiment 2, the words were presented in isolation, and so additional acoustic information provided in lengthened suffixes helped speed reaction times. In Experiment 3, however, the words were presented in a sentence context. Since a third-person singular *–s* suffix must be licensed by a third-person singular verb, participants

[1] We are grateful to a reviewer for this suggestion.

would have known which form of the verb to expect by the time they heard the target. The additional acoustic information in a lengthened suffix would therefore have been of little use in confirming their identification of the target word, and so listeners did not attend to it.

The third result is the enormous advantage for both accuracy and reaction time of hearing stimuli in a use context compared to a mention context. This advantage provides indirect support for the cognitive return-on-investment account. In a use context, the target words are integrated into the sentence context in such a way that surrounding semantic and syntactic cues can help listeners predict an upcoming word, and confirm upon hearing it that the acoustic signal in the target position did in fact match the prediction. However, it is not straightforward that this is the sole reason for the advantage for the words in the use context, because a post-hoc analysis of fillers revealed a similar advantage of use context for filler items, both words and—crucially—non-words. If the facilitation in use context sentences springs from the fact that semantic and syntactic context provides useful cues in predicting and identifying real words, then the facilitation of the use context for non-words, which are impossible to predict, cannot possibly be the result of the same process. It could, perhaps, result from the fact that clear cues and predictions provided by the use context make it easier for participants to identify *violations* of those predictions by the non-words, and give the correct ``non-word'' response. Alternatively, and perhaps more likely, something about the prosodic structure of the use context sentences may have made it easier to identify the target item, and this prosodic advantage affected both real words and non-words in the same way.

General Discussion

In four experiments of increasing task complexity, we tested how listeners draw on

distributional properties of suffixed English verbs during speech perception, and whether they attend to mismatches between the auditory signal and the expected phonetic realizations of those verbs. The lack of interaction between suffix length and paradigmatic probability in Experiment 1b and 2, and between length and context in Experiment 3, suggests that listeners do not seem to attend to mismatches between suffix durations that they hear, and suffix duration patterns that they might expect to hear based on the paradigmatic or contextual probability of the target words. Rather, their speech perception benefits more generally from increased probability, but, crucially, only in some listening contexts. When their attention is directed to single-word stimuli, with a task focusing on sub-lexical constituents, morphologically-derived probability aids perception, as in Experiments 1a-1b. When the task requires more holistic word perception, as in Experiment 2, morphological distributional properties no longer affect perception, but general lexical frequency does. And when the task involves higher-level sentence processing, as in Experiment 3, even effects of lexical frequency weaken or disappear, though effects of sentence context remain.

This pattern of results is striking largely because of the *absence* of predicted effects. Before turning to the implications of these results for theories of speech perception, therefore, it is first worth addressing alternative explanations for these absences. Consider first the absence of any interaction between length and paradigmatic probability in Experiment 1b. It could be, as we suggest above, that listeners depend more on general probabilistic patterns than on the match between actual and expected phonetic realizations, but it could also be that the length manipulation we used did not actually reflect the expected phonetic realizations in the way it was intended to. Kemps and colleagues

(Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005) observed that it was syllable nuclei and codas—not onsets—that were longer in unsuffixed rather than suffixed words. This suggests that durational cues to word structure can affect sub-parts of the stem in different ways (see also White & Turk, 2010). Since we did not consider the duration of subcomponents within the stem, it could be that our acoustic adjustments didn't actually create the intended paradigmatic enhancement effect.

We think, however, that the absence of an interaction between suffix length and paradigmatic probability can not be attributed to an overly crude acoustic manipulation. First, a key distinction between the work presented here and that of Kemps and colleagues (Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005) is the presence of the suffix. Kemps and colleagues were concerned with how the absence of a suffix affects stem duration. In our case, however, all the critical targets are suffixed; it was only the duration of the suffix that varied, not its presence or absence, and so we would not expect to find the same sorts of stem-internal duration patterns that Kemps and colleagues found. Further, even if such stem-internal duration patterns did apply to our stimuli, there is evidence that listeners do not require such nuanced cues to draw on durational information in perception. Blazej and Cohen-Goldberg (2015), for example, manipulated stem duration uniformly, without distinguishing between onsets, nuclei and codas. They nevertheless found that listeners were sensitive to stem duration as a cue for upcoming suffixes in the same way that Kemps and colleagues' listeners were. For these reasons, we think that our acoustic manipulations would have been sufficient to elicit an interaction between stem length and paradigmatic probability, if indeed listeners had been attending to the match between actual and expected pronunciation patterns.

A second absent result worth considering more deeply is the disappearance of the lexical frequency effect in Experiment 3—a particularly striking absence, in light of how robust it is in many psycholinguistic experiments. Therefore, we took a closer look at our lexical frequency measures. In this work we only considered verbal usages of the target word lexemes because we are interested specifically in the effects of the verbal inflectional paradigm. It is possible, however, that the speaker does not make such distinctions, and that the listener therefore has learned associations between pronunciation and general frequency measures, rather than part-of-speech specific measures. We therefore ran a set of post-hoc analysis for all experiments reported here, in which the frequency of non-verbal usages of the target words—a 'residual' frequency, as it were—was added to the model as a covariate. In none of these additional analyses did the inclusion of residual frequency reach significance if the effect of lexical frequency was not significant to begin with.[2] In other words, the absence of a lexical frequency effect in Experiment 3 cannot be attributed to our use of verb-specific frequency measures.

Therefore, although it is never wise to draw too firm a conclusion from negative results and disappearing effects, the results presented here are consistent with a view of speech perception in which flexibility is key. Listeners are capable of drawing on many sources of linguistic information, but they do not actually make use of this information in all situations. This view also appears in the work of Mattys and colleagues. They show that

[2] In some cases the residual frequency term did not reach significance at all; in other cases neither the verb-specific frequency term nor the residual frequency term reached significance if both were included in the same model. In one case, the residual frequency term reached significance and the verb-specific frequency term became insignificant. We believe that these results simply reflect the high degree of correlation between the lexical frequency and residual frequency terms, which effectively competed with each other to explain the same variability in the data.

increasing cognitive load during categorical perception, word segmentation, or phoneme restoration tasks increases participants' reliance on lexical-semantic properties (Mattys, Barden, & Samuel, 2014; Mattys, Brooks, & Cooke, 2009; Mattys & Wiget, 2011). The authors attribute this finding to reduced perceptual acuity towards auditory information. In other words, when the cognitive load interferes with participants' ability to make use of acoustic cues in perceptual tasks, they fall back on the information that they do still have access to—namely, lexical-semantic properties. This account is consistent with the results of Experiment 3 reported here. The RTs in Experiment 3 were more than twice as long as RTs in Experiment 2—indicating that the task of responding to the target word from within a sentence frame was more demanding than the task of responding to a single-word stimulus—and it is in exactly this experiment that participants ceased to pay attention to suffix length (acoustic information) while still attending to the sentence context (lexical-semantic information).

Having a flexible perceptual system means that, depending on the nature of the task, and the demands put upon the listener's attention and processing resources, different types of linguistic information have more or less influence upon speech perception processes, because these details are more or less useful to perception. Paradigmatic probability and its relationship to suffix duration is a systematic pattern, but incorporating it into word identification may provide only minimal improvement in word identification. After all, to draw on paradigmatic probability in the first place, the listener must know which morphological paradigm is at issue, and to do that the verb stem must be identified. By the time the verb stem is known, however, the search space of word identity has been massively narrowed down, such that for nouns only two possible forms remain in English,

and for verbs no more than five. When the processing resources of the listener have nothing better to do, and when they are already directed to pay attention to the phonetic realization of the suffix by the task demands, as in Experiments 1a-b, then paradigmatic probability can aid in the phoneme monitoring task. But this aid is minimal, and not worth the resource expenditure when the tasks become more complex.

Unlike paradigmatic probability, lexical frequency can begin to narrow down the search space from the very beginning of the target word onset, and for that reason it is vastly more useful in word identification that paradigmatic probability. This improved utility shows up in many different perceptual experiments. Words with higher lexical frequency are processed more rapidly in spoken sentence comprehension (Ferreira, Henderson, Anes, Weeks, & McFarlane, 1996), for example, as well as single-word tasks of the sort reported here, and people rely on it to identify disambiguate unclear pronunciations (Connine, Titone, & Wang, 1993). Since lexical frequency is so useful in speech perception, it makes sense that people would continue to make use of it in all but the most complex processing tasks. Only when participants are required to identify a target word from a surrounding sentence and decide whether it is a real word of English do the processing resources otherwise devoted to making use of word frequency patterns become redirected, and frequency effects disappear.

Conclusion

The work presented here offers another facet of the ongoing work investigating flexibility of speech perception. Although paradigmatic probability has reliable effects on speech production, and although listeners are sufficiently sensitive to paradigmatic probability in ideal listening conditions for their processing speed to be affected, those effects disappear

in tasks that more closely approximate listening naturalistic speech. The resulting picture is of a speech perception system that has the potential to identify and employ many more patterns in the speech stream than it actually uses in most contexts.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(38), 419–439. https://doi.org/10.1006/jmla.1997.2558

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

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

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5 Pt 1), 3048–3058. https://doi.org/10.1121/1.2188331

Baayen, R. H., Levelt, W. J. M., Schreuder, R., & Ernestus, M. T. C. (2008). Paradigmatic structure in speech production. *Proceedings of the Chicago Linguistics Society*, *43*, 1–29.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows:

Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon*, *2*(3), 419–463. https://doi.org/10.1075/ml.2.3.06baa

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. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using {lme4}. *Journal Of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *Journal of the Acoustical Society of America*, *133*(4), 2350–2366. https://doi.org/10.1121/1.4794366

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

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

Blazej, L. J., & Cohen-Goldberg, A. M. (2015). Can We Hear Morphological Complexity Before Words Are Complex ? *Journal of Experimental Psychology. Human Perception*

and Performance, *41*(1), 50–68.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer. Retrieved from http://www.praat.org

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*, 991-- 997. https://doi.org/10.3758/s13428-012-0190-4

Cohen, C. (2014). Probabilistic reduction and probabilistic enhancement. *Morphology*, *24*(4), 291–323. https://doi.org/10.1007/s11525-014-9243-y

Cohen, C. (2015). Context and paradigms: Two patterns of probabilistic pronunciation variation in Russian agreement suffixes. *The Mental Lexicon*, *10*(3), 313–338.

Connine, C. M., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/0278-7393.19.1.81

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465–480. https://doi.org/10.1006/jmla.1999.2688

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 218–244. https://doi.org/10.1037/0096-1523.28.1.218

Ferreira, F., Henderson, J. M., Anes, M. D., Weeks, P. A., & McFarlane, D. K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving-window technique. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *22*(2), 324–335. https://doi.org/10.1037/0278-7393.22.2.324

Gahl, S., & Garnsey, S. M. (2004). Knowledge of Grammar, Knowledge of Usage: Syntactic Probabilities Affect Pronunciation Variation. *Language*, *80*(4), 748–775. https://doi.org/10.1353/lan.2004.0185

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 Linguistics Society* (Vol. 35, pp. 151–166).

Heinrich, A., Flory, Y., & Hawkins, S. (2010). Influence of English r-resonances on intelligibility of speech in noise for native English and German listeners. *Speech Communication*, *52*(11–12), 1038–1055. https://doi.org/10.1016/j.specom.2010.09.009

Hyönä, J., Vainio, S., & Laine, M. (2002). A morphological effect obtains for isolated words but not for words in sentence context. *European Journal of Cognitive Psychology*, *14*(4), 417–433. https://doi.org/10.1080/09541440143000131

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. *Frequency and the Emergence of Linguistic Structure*, 229–254.

Kemps, R. J. J. K., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005). Prosodic cues for morphological complexity: the case of Dutch plural nouns. *Memory & Cognition*, *33*(3), 430–446. https://doi.org/10.3758/BF03193061

Kemps, R. J. J. K., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, *20*(1/2), 43–73. https://doi.org/10.3758/BF03193061

Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588–611. https://doi.org/10.1016/j.jml.2012.04.003

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America*, *121*(4), 2261–2271. https://doi.org/10.1121/1.2537393

Lehiste, I. (1972). {T}he timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America*, *51*(6.2), 2018–2024. https://doi.org/10.1121/1.1913062

Lukyanenko, C., & Fisher, C. (2016). Where are the cookies? Two- and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition*, *146*, 349–370. https://doi.org/10.1016/j.cognition.2015.10.012

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156. https://doi.org/10.1080/03640210709336987

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. https://doi.org/10.3758/s13428-011-0168-7

Mattys, S. L., Barden, K., & Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin & Review*, *21*(3), 748–754. https://doi.org/10.3758/s13423-013-0544-7

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*(3),

203–243. https://doi.org/10.1016/j.cogpsych.2009.04.001

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*(2), 145–160. https://doi.org/10.1016/j.jml.2011.04.004

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, *47*(8), 908–930. https://doi.org/10.1016/j.cortex.2011.02.019

Moscoso Del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*(1), 1–18. https://doi.org/10.1016/j.cognition.2003.10.015

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*, 739–773.

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51–89. https://doi.org/10.1016/S0010-0277(03)00139-2

Schuppler, B., Van Dommelen, W. A., Koreman, J., & Ernestus, M. (2012). How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics*, *40*(4), 595–607.

https://doi.org/10.1016/j.wocn.2012.05.004

Tabak, W., Schreuder, R., & Baayen, R. H. (2005). Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in Dutch. *Linguistic evidence—Empirical, Theoretical, and Computational Perspectives*, (1993), 529–555.

Tabak, W., Schreuder, R., & Baayen, R. H. (2010). Producing inflected verbs: A picture naming study. *The Mental Lexicon*, *5*(1), 22–46. https://doi.org/10.1075/ml.5.1.02tab

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, *1*(2), 147–165. https://doi.org/10.1515/LANGCOG.2009.008

Tily, H., & Kuperman, V. (2012). {R}ational phonological lengthening in spoken {D}utch. *Journal of the Acoustic Society of America*, *132*(6), 3935–3940.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. https://doi.org/10.1037//0278-7393.25.2.394

White, L., & Turk, A. E. (2010). English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics*, *38*(3), 459–471. https://doi.org/10.1016/j.wocn.2010.05.002

Table 1
Correlations between numeric variables used in Experiment 1a. No correlations exceeded .1 in absolute value.

	Log previous RT	Trial number	Log frequency	Paradigmatic probability
Stem duration	-.054	.037	.037	-.070
Log previous RT		-.054	.026	-.002
Trial number			.014	.013
Log frequency				.091

Table 2

a) *Regression coefficients, standard errors, and t-values for the final RT model from Experiment 1a. Random effects included intercepts for subject and word.*

Control model	*β*	*SE(β)*	*t*
Intercept (length=norm)	5.522	0.093	59.56
stem duration (ms)	0.0008	0.0001	13.59
log previous RT	0.156	0.013	12.34
trial number	-0.0001	0.00002	-5.20
log frequency	-0.006	0.002	-3.52
length			
length = short	-0.010	0.006	-1.81
length = long	0.001	0.006	3.92
paradigmatic probability * length			
length = norm	0.008	0.006	1.47
length = short	-0.015	0.006	-2.32
length = long	-0.017	0.006	-2.65

b) *Results of a log-likelihood ratio test testing whether each predictor of interest significantly improves model fit.*

Predictor	χ^2 (*df*)	*p*-value
length (added to model containing paradigmatic prob.)	33.59 (2)	< 0.001
paradigmatic probability (added to model containing length)	0.28 (1)	0.595
paradigmatic probability * length (added to model containing length only)	8.64 (3)	< 0.05

Table 3

Correlations between numeric variables used in Experiment 1b. No correlations exceeded .14 in absolute value.

	Log previous RT	Trial number	Log frequency	Paradigmatic probability
Stem duration	.003	-.005	-.136	-.054
Log previous RT		-.091	-.006	.008
Trial number			-.008	-.010
Log frequency				.125

Table 4

Regression coefficients, standard errors, and t-values for the replication RT model from Experiment 1b. Random effects included intercept and slopes for paradigmatic probability and suffix length by subject, and intercept and slope for suffix length by item.

	β	*SE(β)*	*t*
Intercept (length=norm)	5.48	0.056	98.70
stem duration (ms)	0.001	0.00005	23.00
log previous RT	0.118	0.007	16.00
trial number	-0.00003	0.00001	-3.35
log frequency	-0.003	0.001	-3.20
length			
length = short	-0.007	0.004	-1.97
length = long	0.002	0.002	0.51
paradigmatic probability	-0.006	0.002	-2.64

Table 5

Correlations between numeric variables used in Experiment 2. No correlations exceeded .1 in absolute value. Since the different distribution of discarded observations results in a slightly different balance of individual items in the data set, the correlation between frequency and paradigmatic probability does not exactly match that of Experiment 1a, even thought the target words were identical.

	Previous RT	Trial number	Log frequency	Paradigmatic probability
Stem duration	-.004	.024	-.009	-.026
Previous RT		-.062	-.015	-.008
Trial number			.016	.015
Log frequency				.092

Table 6

a) *Regression coefficients, standard errors, and t-values for the final RT model from Experiment 2. Random effects included intercepts for subject and word.*

Control model	*β*	*SE(β)*	*t*
Intercept (length=short)	322.93	47.77	6.76
stem duration (ms)	1.04	0.08	13.55
log previous RT	0.11	0.01	12.00
trial number	-0.18	0.02	-8.65
log frequency	-5.58	2.00	-2.79
length			
length = norm	-24.65	6.33	-3.85
length = long	-30.70	7.98	-3.85
paradigmatic probability * length			
length = short	-2.92	6.49	-0.45
length = norm	1.64	6.48	0.25
length = long	.010	6.47	0.00

b) Results of a log-likelihood ratio test testing whether each predictor of interest significantly improves model fit.

Predictor	χ^2 (*df*)	*p*-value
length (added to model containing paradigmatic prob.)	17.80(2)	< 0.001
paradigmatic probability (added to model containing length)	0.20(1)	0.66
paradigmatic probability * length (added to model containing neither)	0.28(3)	0.96

Table 7

a) *Regression coefficients, standard errors, and z-values for the accuracy models from Experiment 3. Interactions did not improve model fit, and parameter estimates for the interactions are therefore not shown. Random effects included intercepts for word and subject, and slopes for lexical frequency by subject.*

	β	*SE(β)*	*z*	*p(z)*
Intercept (length = long, context = mention)	2.09	0.32	6.52	< 0.001
log frequency	0.10	0.05	1.97	< 0.05
context				
context = use	1.58	0.21	7.55	< 0.001
length				
length = norm	0.45	0.22	2.02	< 0.05
length = short	-0.07	0.20	-0.34	0.73
paradigmatic probability	-0.14	0.13	-1.10	0.27

b) *Results of a log-likelihood ratio test testing whether each predictor of interest significantly improves model fit. In each row, the model comparison compares a model containing each predictor to a model containing everything in Table 7a* except *the predictor in that row.*

Predictor	χ^2 (*df*)	*p*-value
Log frequency	3.55 (1)	0.06
Context	66.9 (1)	< 0.001
Length	6.15 (2)	< 0.05
Paradigmatic probability	1.13 (1)	0.29

Table 8
Correlations between numeric variables used in Experiment 3. No correlations exceeded .16 in absolute value.

	Previous RT	Trial number	Log frequency	Paradigmatic probability
Stem duration	-.010	.025	-.15	-.13
Previous RT		-.115	.012	-.042
Trial number			.003	.019
Log frequency				-.018

Table 9

a) *Regression coefficients, standard errors, and z-values for the reaction time model from Experiment 3. Interactions did not improve model fit, and parameter estimates for the interactions are therefore not shown. Random effects included intercepts for subject and sentence, and slopes for context by subject.*

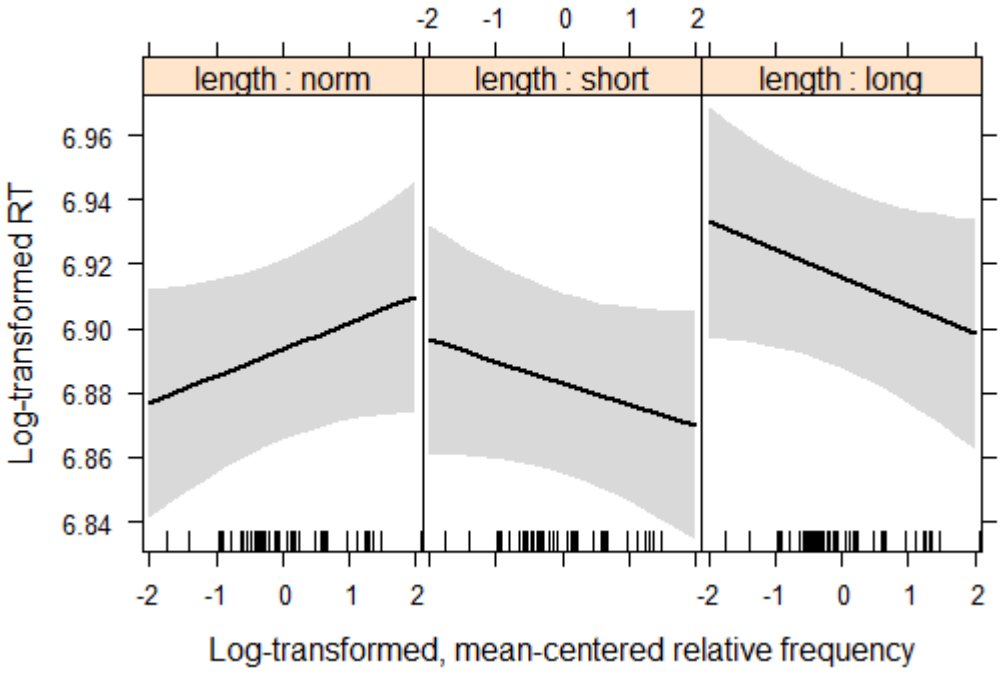
	β	*SE(β)*	*t*
Intercept (length = norm, context = mention)	873.11	43.14	20.74
word duration (ms)	0.41	0.08	5.29
trial number	-0.44	0.05	-8.81
previous RT	0.08	0.01	7.72
log frequency	-1.77	1.92	-0.92
context			
context = use	-125.46	10.90	-11.51
length			
length = short	4.80	8.74	0.55
length = long	6.74	8.73	0.77
paradigmatic probability	-1.55	5.58	-0.28

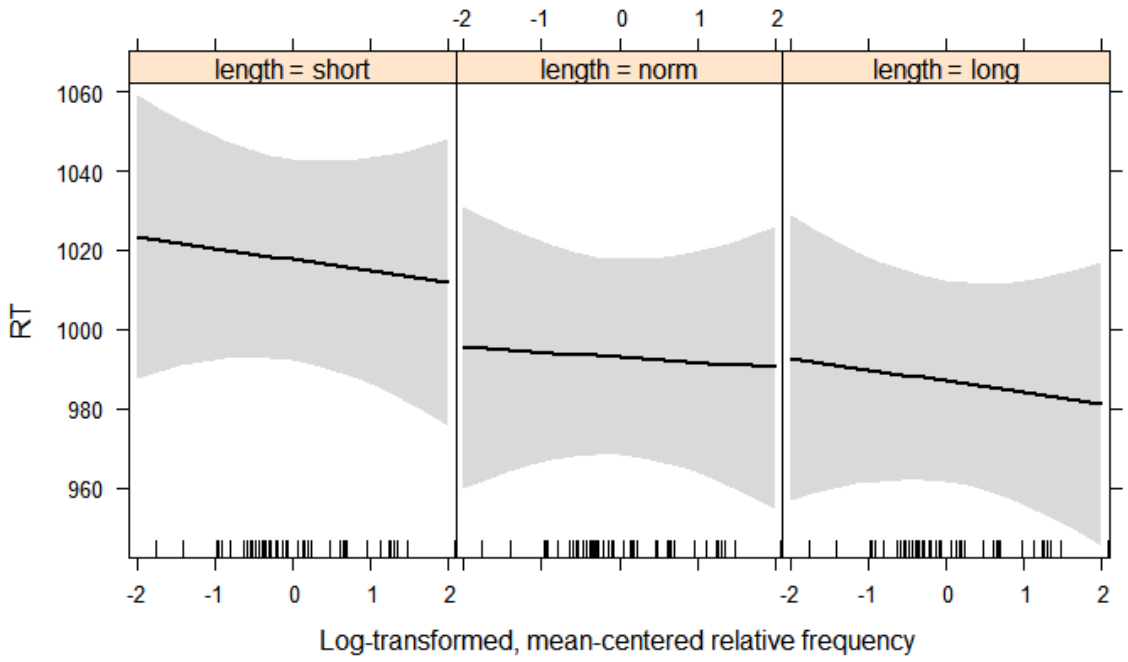
b) *Results of a log-likelihood ratio test testing whether each predictor of interest significantly improves model fit. In each row, the model comparison compares a model containing each predictor to a model containing everything in Table 8a* except *the predictor in that row.*

Predictor	χ^2 (*df*)	*p*-value
Log frequency	0.85 (1)	0.36
Context	69.45 (1)	< 0.001
Length	0.69 (2)	0.71
Paradigmatic probability	0.08(1)	0.78

Figure 1: Partial effects plot showing the interaction between length and paradigmatic probability in reaction times from Experiment 1.

Figure 2: Partial effects plot showing the effects of length and paradigmatic probability in reaction times from Experiment 2. Participants responded more quickly to norm and long suffixes (right two panels) than to short suffixes (left panel).


length : norm
length : short
length : long
Log-transformed RT
6.96
6.94
6.92
6.90
6.88
6.86
6.84
-2
-1
0
1
2
Log-transformed, mean-centered relative frequency


length = short
length = norm
length = long
RT
1060
1040
1020
1000
980
960
-2
-1
0
1
2
Log-transformed, mean-centered relative frequency