Journal of Experimental Psychology: Learning, Memory, and Cognition

Changes in Informativity of Sentential Context Affects Its Integration With Subcategorical Information About Preceding Speech

Wednesday Bushong and T. Florian Jaeger

Online First Publication, February 13, 2025. https://dx.doi.org/10.1037/xlm0001443

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Learning, Memory, and Cognition

ISSN: 0278-7393

https://doi.org/10.1037/xlm0001443

Changes in Informativity of Sentential Context Affects Its Integration With Subcategorical Information About Preceding Speech

Wednesday Bushong[1, 2, 3] and T. Florian Jaeger[4, 5]

[1] Department of Psychology, University of Hartford
[2] Program in Cognitive and Linguistic Sciences, Wellesley College
[3] Department of Psychology, Wellesley College
[4] Department of Brain and Cognitive Sciences, University of Rochester
[5] Goergen Institute for Data Science and Artificial Intelligence, University of Rochester

Spoken language understanding requires the integration of incoming speech with representations of the preceding context. How rich the information is that listeners maintain in these contextual representations has been a long-standing question. Under one view, subcategorical information about the preceding input—including any uncertainty about the underlying categories—is quickly discarded due to memory limitations. Alternative views hold that listeners maintain some subcategorical information far beyond word boundaries. This would facilitate more effective integration with subsequent context, under the assumption that subsequent context is informative about the preceding input. We thus ask whether listeners are sensitive to changes in the informativity of subsequent context by changing the expected utility of subcategorical information maintenance. In three experiments, we manipulate how informative subsequent context is about words that occur six to nine syllables earlier. We find that reduced informativity leads listeners to down-weight the importance of subsequent context. This suggests that listeners can adjust the degree to which they maintain subcategorical information. We do, however, also identify alternative interpretations that affect not only the present results but also the interpretation of previous work on subcategorical information maintenance.

Keywords: speech perception, information maintenance, right context, cue integration

Supplemental materials: https://doi.org/10.1037/xlm0001443.supp

During spoken language understanding, listeners integrate the incoming acoustic signal with contextual representations derived from the preceding speech input. It is now clear that these contextual representations are surprisingly rich, including a wide range of lexical, discourse, and talker information. Questions remain, however, about the extent to which listeners maintain even richer *subcategorical information* about recently processed speech as part of these contextual representations.

According to some early accounts of spoken word recognition (described in more detail below), listeners categorize incoming speech segment by segment. Simultaneous with each categorization, listeners were thought to discard all perceptual or phonetic information about the preceding input, including any uncertainty about the recognized categories. Under such accounts, the input to spoken word recognition was sequences of recognized phonological segments, devoid of any subcategorical information.[1] Later work repeatedly revised these

[1] Following our previous work, we use the term *subcategorical information* as an umbrella term for any type of information beyond discrete category identity. This includes uncertainty about category identity or—equivalently for the present purpose—gradient activation of alternative category candidates (contrary to Caplan et al., 2021; see, e.g., Bicknell et al., 2025; Burchill et al., 2018, pp. 19–20; Bushong, 2020, pp. 2–4, 148–149) but could theoretically include richer information, closer to phonetic or even perceptual representations. Such an umbrella term is necessary, as the nature of maintained subcategorical information is still under debate (Burchill et al., 2018; Caplan et al., 2021; for review, see Burchill, 2023).

L. Robert Slevc served as action editor.

Wednesday Bushong https://orcid.org/0000-0002-1837-0689

The results of Experiment 1 were presented at the 24th Annual Conference on Architectures and Mechanisms for Language Processing. The results of Experiments 1 and 2 were presented at the 41st Annual Conference of the Cognitive Science Society and Wednesday Bushong's PhD dissertation.

This research was partially funded by the National Institute of Child Health and Human Development (Grant R01 HD075797 awarded to T. Florian Jaeger) and an National Science Foundation Research Traineeship (Grant #1449828 awarded to Wednesday Bushong). The views expressed here do not necessarily reflect those of the funding agencies. The authors thank Robbie Jacobs for discussions about cue integration, as well as Evan Hamaguchi and Chelsea Marsh for assistance with stimulus sentence creation and recording.

Wednesday Bushong played a lead role in conceptualization, formal analysis, methodology, visualization, writing–original draft, and writing–review and editing. T. Florian Jaeger played a lead role in formal analysis, funding acquisition and a supporting role in visualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Wednesday Bushong, Department of Psychology, Wellesley College, 106 Central Street, Wellesley, MA 02481, United States. Email: wednesday.bushong@wellesley.edu

assumptions, suggesting that some subcategorical information—at least uncertainty about the recognized words or their referents—is maintained in short-term memory for substantially longer than previously thought, potentially for dozens of syllables (e.g., Brown-Schmidt & Toscano, 2017; Burchill et al., 2018; Connine et al., 1991; Falandays et al., 2020; Gwilliams et al., 2018; McMurray et al., 2009; Szostak & Pitt, 2013; for review, see Bicknell et al., 2025; Dahan, 2010).

In the present work, we set out to better understand the memory processes involved in such maintenance. We ask whether listeners strategically modulate the maintenance and integration of subcategorical information depending on its expected utility, rather than passively maintaining subcategorical information at all times. We begin by introducing the necessary background to motivate the present work. Then, we briefly anticipate issues with the logic of our design that affect its interpretation. This includes assumptions that, we believe, pertain to *all* research on subcategorical information maintenance, affecting what can be concluded from existing results in the field.

Evidence That Listeners Maintain Subcategorical Information

There is now ample evidence that listeners' *long-term* representations of speech seem to contain fine-grained subcategorical detail (e.g., Hanulikova, 2022; Hay et al., 2019; Johnson et al., 1999; Niedzielski, 1999; Walker & Hay, 2011; for review, see Foulkes & Hay, 2015), in line with exemplar (Johnson, 1997) and episodic theories of speech perception (Goldinger, 1998). This leaves open, however, for how long subcategorical information about the most recently processed speech input remains available in *short*-term memory resources. A priori, there are reasons to expect listeners to maintain subcategorical information. Consider, for example, that phonetic cues to the identity of phonological categories are often temporally distributed across the speech signal. For instance, cues to syllable-final stop voicing in English (e.g., "tap" vs. "tab") include both the duration of the preceding vowel and the closure of the stop itself (Klatt, 1976). Spoken word recognition would thus be facilitated if listeners retain relevant subcategorical information about the preceding vowel in memory. Indeed, *optimal* integration of phonetic cues that are distributed across the speech signal is only possible if subcategorical information—at least uncertainty about the underlying categories—is still available when the later cue is processed (Bicknell et al., 2025; Bushong, 2020).

On the other hand, human perceptual memory is finite, and the amount of information contained in the speech signal every second is substantial—orders of magnitude larger than the amount of information required to encode, for instance, only the sequence of recognized segments. Classic studies initially seemed to suggest that auditory memory decays within fractions of a second (for review and critique, see Crowder, 1982). One early estimate, for example, suggested that rich auditory information decays within 200–300 ms (about the duration of a syllable in conversational speech), though—in the absence of intervening input—some phonetic information seemed to be maintained for 1–3 s (Cowan, 1984). Though later work substantially revised these estimates upward (e.g., up to 10 s for auditory information in the absence of intervening material; Atienza et al., 2000; Bottcher-Gandor & Ullsperger, 1992), the intuition that strong memory pressures force listeners to rapidly discard subcategorical information has remained influential in some parts of the field (for review, see Christiansen & Chater, 2016).

Models of spoken word recognition, too, initially made strong assumptions about memory pressures that were increasingly relaxed over subsequent decades. One influential early model of spoken word recognition, the cohort model (Marslen-Wilson & Welsh, 1978), held that listeners recognized speech segment by segment, discarding information about recognized segments. This strong hypothesis has since been soundly rejected by both behavioral (e.g., Allopenna et al., 1998; McMurray et al., 2009) and brain imaging evidence (Gwilliams et al., 2018, 2022) that subcategorical information about earlier segments remains available for integration with subsequent input at least for the duration of a word. For example, McMurray and colleagues manipulated the primary cue to English word-initial stop voicing (voice onset time or VOT), creating word–nonword continua like "[b/p]arricade." In these stimuli, the subsequent lexical context effectively disambiguates the preceding input. Critically, eye-movement data revealed that the processing of the later disambiguating input interacted with the gradient phonetic realization—the VOT—of the [b/p] sound. Listeners thus maintained some gradient subcategorical information about VOT until the end of the word and integrated it with the subsequent disambiguating context.

Later models of spoken word recognition relaxed or completely abandoned the recognize-and-discard idea of the cohort model (e.g., Luce & Pisoni, 1998; Magnuson et al., 2020; McClelland & Elman, 1986; Norris & McQueen, 2008; Oden & Massaro, 1978). These theoretical developments are supported by findings that some subcategorical information remains available for integration with subsequent context beyond word boundaries (e.g., Brown-Schmidt & Toscano, 2017; Bushong & Jaeger, 2019a; Liu & Jaeger, 2018; Szostak & Pitt, 2013), with some studies detecting gradient effects of preceding input as far as 35 syllables downstream (the longest distance tested; Falandays et al., 2020).

The Present Work

Here, we ask whether the extent to which subcategorical information is maintained depends on the expected utility of such maintenance. The statistics of natural language are such that subsequent context *typically* is informative (e.g., Aylett & Turk, 2004; Bell et al., 2009; Qian & Jaeger, 2012), so that maintaining subcategorical information about preceding speech input for integration with subsequent context would facilitate comprehension (Bicknell et al., 2025). However, the degree to which subsequent context is informative about preceding input does vary depending on, for example, the types of contexts a word tends to occur in (Bell et al., 2009). This raises the possibility that listeners maintain subcategorical information to different extents, depending on the expected informativity of subsequent context (or, in the terminology of rational resource-use accounts of cognition, depending on the *expected utility* of information maintenance; Griffiths et al., 2015; Lewis et al., 2014; Lieder & Griffiths, 2019). Under this view, maintenance of subcategorical information is not solely determined by hard constraints imposed by limits on perceptual or other short-term memory. Rather, information maintenance is seen as a resource-consuming process that listeners implicitly aim to use only to the extent that it is expected to facilitate speech perception.

To assess whether this view has potential merit, we present three experiments on spoken word recognition that investigate the relationship between the informativity of subsequent context and the maintenance of subcategorical information. All three experiments use an initial exposure phase to manipulate, between participants, the availability of informative subsequent context. One group of participants hears sentences that always contain subsequent contextual information that is *informative* for categorization, whereas the other group of participants hears sentences with comparatively *uninformative* subsequent context. Following exposure, we then test whether participants in each group continue to maintain subcategorical representations about the target word beyond the word boundary.

Beyond our primary question, we also examine whether maintenance of subcategorical information is the "default" in spoken word recognition, rather than a response to the specific task demands of an experiment. Given that subsequent context, on average, *is* informative about preceding speech input, maintenance would be expected to be the typical behavior under the view we are testing here. We address this question in two ways: by asking whether subcategorical information is maintained even during the first trial of the experiment (see also Bushong & Jaeger, 2019a) and by comparing the extent to which different participants maintain subcategorical information.

Assumptions

Before we describe our experiments, we lay out the logic that has been used in the interpretation of previous work on subcategorical information maintenance. This highlights assumptions that we adopt following previous work and that we revisit in the general discussion.

Our paradigm—in particular, the test phase—closely follows the seminal study by Connine et al. (1991) on the integration of preceding acoustic cues with subsequent ("right") context. This paradigm, described in more detail later, uses sentence recordings like (1), which manipulate two aspects of the recorded speech. First, the VOT of the target word (marked by "?ent") is manipulated to range from values that bias toward "dent" to values that bias toward "tent." Second, the subsequent context is manipulated to bias either toward "tent" (1a) or "dent" (1b). After each recording, participants respond whether they heard "tent" or "dent."

1a. *Tent*-biasing context: After the ?ent Sue had found in the campgrounds collapsed, we went to a hotel.

1b. *Dent*-biasing context: After the ?ent Sue had found in the teapot was noticed, we threw it away.

We interpret our results under the same assumptions introduced by Connine et al. (1991) and—often implicitly—adopted in subsequent work (e.g., Bicknell et al., 2025; Szostak & Pitt, 2013): If participants' responses are gradiently affected by both the subcategorical acoustic cues (VOT) and subsequent context, this constitutes evidence that participants maintained subcategorical information about the initial input. If listeners instead immediately categorize the stimulus based on VOT, then discard VOT, and never revise their categorization, this would result in effects of only VOT. Vice versa, effects of only subsequent context would mean that listeners either ignored the preceding subcategorical information or allowed subsequent context to completely override any effects of that information. Effects of only VOT or only context thus would not constitute evidence for subcategorical information maintenance. The same assumptions have been applied to the interpretation of, for example, eye-movement behavior in visual world experiments on subcategorical information maintenance: If participants' eye movements during, or following, the processing of the subsequent context are affected by both the preceding subcategorical information and the context, then participants maintained subcategorical information (e.g., Brown-Schmidt & Toscano, 2017; Falandays et al., 2020).

Under these assumptions, *reduced* maintenance of subcategorical information due to reduced informativity of subsequent context could show in a number of different ways. First, listeners might abandon maintenance altogether on some or all trials, categorize the *?ent* stimulus based on the VOT, and then discard any memory of that VOT. This strategy would require no maintenance of subcategorical information and no integration with subsequent context. This would reduce or completely remove effects of subsequent context on those trials, without changes in the effects of VOT (compared to the effects prior to the reduction of maintenance). This hypothesis closely follows the assumptions of previous work, and it is the hypothesis that we had in mind when developing our design. Second, listeners might instead categorize the *?ent* stimulus based on VOT and discard VOT (as in the first possibility), but then change their categorizations based on subsequent context. Like the first possibility, this strategy, which we have previously referred to as *categorize-discard-and-switch* (Bushong & Jaeger, 2019c), would require no maintenance. Unlike the first possibility, however, this strategy would result in reduced effects of VOT, in addition to reduced effects of context.[2] Critically, all of these possibilities are expected to result in reduced effects of subsequent context. This is thus the core prediction that we set out to test. Additionally, the presence or absence of reduced effects of VOT would shed light on *how* listeners modulate the maintenance of subcategorical information.[3]

There are, however, plausible alternative explanations for the signature results that have previously been interpreted as evidence for subcategorical information maintenance. In the general discussion, we identify several such alternatives. One of them arises when one abandons the assumption that all participants exhibit the same strategies on all trials. Imagine, for example, that participants sometimes respond based on only the VOT and sometimes based on only the subsequent context. Compared to maintenance of subcategorical information, this would result in smaller effects of VOT and context (since each effect only occurs on a subset of the trials, being reduced or zero on all other trials). It would, however, predict the same qualitative pattern across trials and participants as subcategorical information maintenance: effects of both VOT and context. The presence of these effects by itself is thus insufficient to conclude subcategorical information maintenance. The present study does *not* resolve this issue, but we describe how future work can distinguish between these very different theoretical explanations.

[2] Specifically, this prediction holds when standard analysis approaches are used that do not consider complex nonlinearities in the effect of VOT and its interactions with context (Bushong & Jaeger, 2019c). We return to this and related points in the General Discussion section.

[3] Conversely, if recently experienced input was to suggest that VOT has become less informative about stop voicing, listeners might reduce maintenance by maintaining *less* or *more noisy* subcategorical information about VOT. The present study does not test this hypothesis.

Open Science Statement

All stimulus recordings, data, and analyses are available via the Open Science Framework at https://osf.io/cypg3 (Bushong & Jaeger, 2024).

Experiment 1

Panel A of Figure 1 illustrates the between-participant design of Experiment 1. Unbeknownst to participants, the experiment consisted of two phases. The initial exposure phase manipulated, between participants, whether subsequent context was informative for the task. The test phase was identical across participants and assessed the consequences of exposure. During test, subsequent context was always informative.

Critically, exposure manipulated the *availability* of informative subsequent context—that is, whether the context following the target word contained lexical context that strongly biased toward "tent" or "dent." We did not manipulate the *reliability* of the contextual cues themselves. That is, we did not manipulate how often a given lexical context co-occurred with particular acoustic cue values.[4] This serves our goal to test whether listeners are sensitive to changes in the expected informativity of subsequent context and whether this can alter the degree to which listeners maintain subcategorical information about the preceding speech input. At the same time, it keeps constant the information listeners gain during test *if they pay attention to subsequent context*. This differs from studies on cue (re)weighting, in which exposure affects the amount of information carried by one or more of the cues (e.g., Jacobs, 2002; Toscano & McMurray, 2010; for review, see Schertz & Clare, 2020; specifically, for the integration of acoustic and lexical cues, Bushong & Jaeger, 2019a; Giovannone & Theodore, 2021). We revisit this point in the General Discussion section.

Method

Experiment 1 was approved by the University of Rochester Research Subjects Review Board. It was not preregistered.

Participants

We recruited 128 native English-speaking participants (64 participants for each of the two between-participant conditions). Participants were recruited in November of 2018 from Amazon Mechanical Turk and rewarded $3.00 for their participation (for a targeted rate of at least $6.00/hr). The targeted number of participants was initially determined based on previous work in similar paradigms with 99% power to detect the effect of subsequent context (Bicknell et al., 2025; Bushong & Jaeger, 2019a). Below, we present post hoc power analyses that assess the power of the present study.

In past work (Bushong & Jaeger, 2017, 2019a), we excluded participants who did not use VOT (more "tent" responses for higher VOTs). Here, we instead followed earlier work and did not exclude any participants (Connine et al., 1991; Szostak & Pitt, 2013). Unlike earlier work, however, we ask whether individual participants exhibit both effects required to conclude the presence of subcategorical information maintenance: effects of VOT *and* effects of subsequent context. This allows us to determine—for the first time, as far as we know—whether maintenance of subcategorical information is the "default" across participants (at least in paradigms like ours).[5]

We collected demographic information per the National Institutes of Health reporting requirements. In Experiment 1, 44.5% of participants were female, 53.2% male, and 2.3% declined to respond. With respect to ethnicity, 3.1% reported as Hispanic, 94.5% as non-Hispanic, and 2.4% declined to respond. With respect to race, 83.5% of participants identified as White, 8.5% as Black or African American, 4.6% as Asian, 1.5% as American Indian/Alaska Native or Native Hawaiian or other Pacific Islander, 1.5% as other, and 3% declined to respond (percentages add up to more than 100% because we counted participants who responded "more than one race" toward each of their self-identified races). The mean age of our participants was 37.05 years (SD = 11.76, interquartile range = [28, 44]; 2% declined to respond). Since we had no specific predictions about demographic effects, we did not include demographic variables in our analyses.

Materials

We constructed 12 sentence items, each forming a triplet like the following (see the Appendix for the full list of stimuli):

2a. *Tent*-biasing context: After the ?ent Sue had found in the campgrounds collapsed, we went to a hotel.

2b. *Dent*-biasing context: After the ?ent Sue had found in the teapot was noticed, we threw it away.

2c. Neutral context: After the ?ent was noticed, we continued on our way.

We manipulated two aspects of the sentence stimuli. First, we acoustically manipulated the "?" to range between /d/ and /t/ by changing the value of its VOT, the primary cue distinguishing voiced from voiceless syllable-initial stop consonants in English. Based on norming and previous work, we chose to test VOT values of 10, 40, 50, 60, 70, and 85 ms to cover a perceptual range from unambiguous /d/ to unambiguous /t/ with ambiguous points in between.

Second, we manipulated whether later context biased toward a /t/-interpretation (2a), /d/-interpretation (2b), or neither (2c). Informative words in the subsequent context—if present—occurred between six and nine syllables after the target word, as in examples (2a) and (2b). The sentences in (2a) and (2b) are highly similar in structure to those of Connine et al. (1991) and follow-up work. In Connine et al. (1991) and other previous work, subsequent context was *always* highly informative about *tent* versus *dent* categorization. Sentence stimuli like (2c) were not employed in previous work.

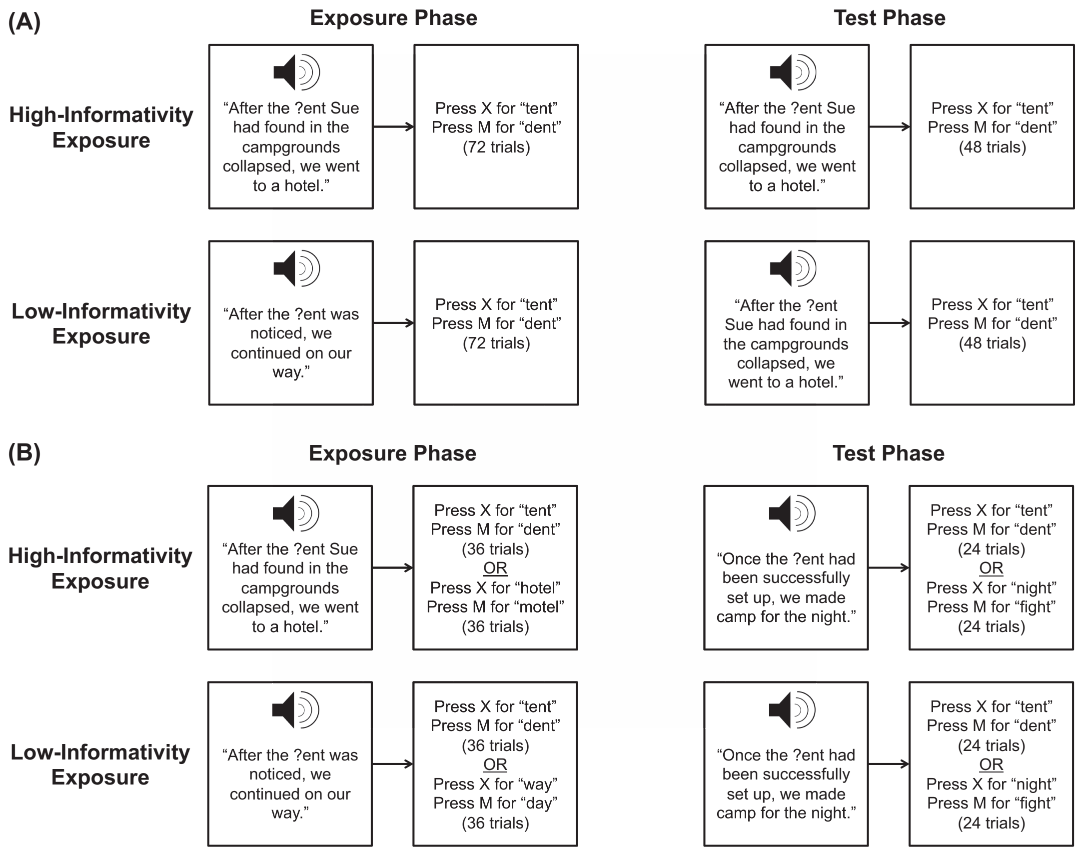
Procedure

Following instructions, participants completed the exposure (72 trials) and test phases (48 trials; see Figure 1). Participants were not informed of this structure, and there were no breaks, additional

[4] Put differently, we manipulated p(subsequent context), rather than p(target = "tent"|subsequent context).

[5] The Supplemental Material (Section 1) presents analyses that follow our previous approach, excluding participants without significant VOT effects. For all three experiments, these analyses replicate the results reported in the main text.

Figure 1
Design of Experiment 1 (A) and Experiments 2–3 (B)



Note. Participants were never informed about the exposure–test structure of the experiment. The task remained constant across the entire experiment, and there was no break or additional instructions between exposure and test. Within each experiment, both exposure groups saw the same test trials. In Experiments 2-3, participants made categorization decisions either about the critical target word or a different word in the sentence; examples of alternate target words are provided in Panel B.

instructions, or change in task between exposure and test: On each trial, participants listened to a sentence recording and responded whether they heard "tent" or "dent" (see Figure 1). This is the same task as in previous work (Bicknell et al., 2025; Bushong & Jaeger, 2019a; Connine et al., 1991; Szostak & Pitt, 2013). Finally, participants completed an exit survey and (optionally) provided National Institutes of Health-requested demographic information.

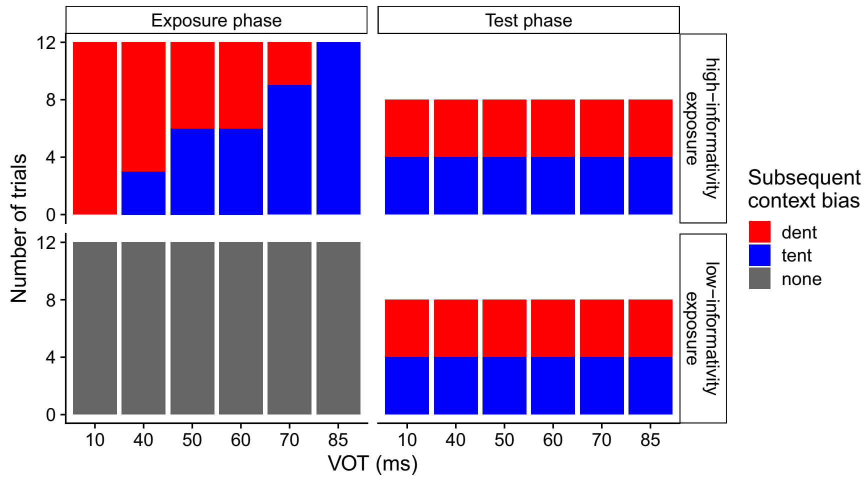
Participants were randomly assigned to one of two exposure groups: low-informativity exposure or high-informativity exposure. During exposure, the low-informativity group only heard sentences with neutral subsequent context (like Sentence 2c above). In those sentences, the only relevant information to categorization was VOT. The high-informativity group, by contrast, always heard sentences that contained informative subsequent context (split evenly between *tent*- and *dent*-biasing contexts), as in previous studies. In the test phase, both groups heard the same sentences. In these sentences, later context was informative (split evenly between *tent*- and *dent*-biasing contexts).

Each participant was randomly assigned to one of 16 experimental lists. These lists were designed so as to (a) keep the experiment reasonably short, (b) employ all 12 items, and (c) counterbalance an additional nuisance factor (whether the items during the test phase differed from those during the exposure phase). In all lists, six items appeared in exposure and four items appeared in test. In half of the lists, the items appearing in test also appeared in exposure (e.g., exposed to Items 1–6 → tested on Items 1–4); in the other half of lists, test items were new (e.g., exposed to Items 1–6 → tested on Items 7–10). Results did not depend on whether items were repeated between exposure and test (see the Discussion section). We also varied which items appeared in exposure: Half of participants were exposed to Items 1–6 in exposure, and the other half to Items 7–12.

For the high-informativity exposure group, all six exposure items were repeated in their two context conditions (*dent*- and *tent*-biasing) and all six VOT steps ($6 \times 2 \times 6 = 72$ total trials). For the low-informativity exposure group, there was only one context condition (see example 2c above), which thus was repeated twice as often. This property of our design—due to the fact that we had only one, rather than two, neutral contexts for each item—was unintended: Participants in the low-informativity condition experienced twice as much sentence repetition during exposure than participants in the high-informativity condition. Conversely, the fact that the test phase never contained neutral contexts (see Figure 2) meant that participants

Figure 2

Distribution of VOTs and Subsequent Biasing Context by Exposure Group and Experiment Phase



Note. Both groups receive identical distributions of VOTs and contexts in the test phase. Note that there are fewer trials in the test phase compared to the exposure phase. VOT = voice onset time. See the online article for the color version of this figure.

in the low-informativity condition never heard exposure sentences repeated during the test phase. We thank Effie Kapnoula for identifying this potential issue, and we address it in Experiment 3.

One final aspect of our design deserves further elaboration. As shown in Figure 2 (left side), our exposure design aimed to avoid—or at least reduce—cue conflicts between VOT and subsequent context. This design decision was motivated by recent evidence that exposure to sentences with cue conflicts—for example, strongly *dent*-biasing VOTs paired with strongly *tent*-biasing subsequent contexts or vice versa—can reduce the effects of subsequent context (Bushong & Jaeger, 2019a). When such cue conflict is avoided, the effects of context and acoustics in listeners' responses remain stable throughout exposure (ibid.), increasing the statistical power to detect differences between exposure groups during test. We thus never paired the /d/-endpoint of the VOT with a *tent*-biasing later context and vice versa; the next two VOT continuum steps from either end (40 ms and 70 ms) occurred with the opposite (conflicting) biasing context 25% of the time; and the two middle VOT steps (50 ms and 60 ms) occurred with both biasing contexts equally (qualitatively following Bushong & Jaeger, 2019a). The resulting VOT distributions were identical for both exposure conditions (cue conflict was trivially avoided for the low-informativity group since subsequent context was never informative about the preceding input).

For the test phase, we follow previous work and fully cross the six VOT steps with the context conditions (Figure 2, right side). This test design allows us to estimate the context effect at all points of the VOT continuum and facilitates comparison to previous work (Bicknell et al., 2025; Connine et al., 1991; Szostak & Pitt, 2013).

Analysis Approach, Predictions, and Power

Analysis Approach

We used R's (R Core Team, 2016) lme4 package (Bates et al., 2014) to fit mixed-effects logistic regressions (Jaeger, 2008), separately to the exposure and test phases. The Supplemental Material (Section 2) presents additional post hoc analysis that analyzes changes in effects across trials, using generalized additive mixed-effects models.

The analysis of the exposure phase tests whether the high-informativity group showed an effect of context during exposure. Since it is impossible to assess context effects during low-informativity exposure, this analysis only used the data from the high-informativity group. The mixed-effects logistic regression predicted participants' /t/-responses as a function of context, VOT, trial, and their interactions. All categorical predictors were sum-coded (context: *tent*-biasing = 0.5, *dent*-biasing = −0.5; group: low-informativity = 0.5, high-informativity = −0.5). VOT was centered and scaled by twice its standard deviation, to ease comparison of effect sizes across predictors (Gelman, 2008). Trial was log-transformed and not centered, so that the coefficients of all other predictors describe the effects on the first trial of the exposure phase (when log-trial is 0). We included the full random effects structure that resulted in successful model convergence: by-subject random intercepts and slopes for context, VOT, and their interaction, and by-item random intercepts and slopes for context (correlations between random effects were not included in this or any other analysis). We predicted significant effects of context and VOT, with no interaction between the two.

The analysis of the test phase tests the three predictions of primary interest to the present study. The regression tested for main effects of context, group, VOT, and trial; all two-way interactions; and the three-way interaction between context, group, and VOT. We excluded the remaining three- and four-way interactions with trial because they were not theoretically relevant (and analyses including them did not converge). Trial was log-transformed and centered to the last trial of the exposure phase, so that the coefficients of all other predictors describe the effects on the first trial of the test phase (when log-trial is 0). As with the exposure phase analysis, each model included the full random effects structure that resulted in successful convergence of the model. For Experiment 1, this was a random intercept plus random slopes for context, VOT, and trial by participant and a random intercept plus random slopes for context, trial, and group by item.

Predictions

We test the hypothesis that listeners can adapt their use of subsequent context based on the current environment. If participants are capable of maintaining subcategorical information about the target word *?ent* until at least the subsequent context word, listeners' responses should be affected by both VOT and context (Bicknell et al., 2025; Connine et al., 1991; Szostak & Pitt, 2013). We thus predicted significant effects of context and VOT. If such subcategorical information maintenance is the default, we should observe these effects in the high-informativity group from the beginning of the exposure phase.

Critically, if listeners monitor the informativity of subsequent context for the target word, this should affect their expectations about the utility of maintaining subcategorical information about the speech input. If these expectations affect the degree to which listeners maintain subcategorical representations, we should find a smaller (or even null) effect of context in the low-informativity exposure group, compared to the high-informativity exposure group. We thus predicted a Context × Group interaction, such that the effect of subsequent context is larger in the high-informativity group compared to the low-informativity group.

Unrelated to the primary goals of the present study, we expected the lack of an interaction between context and VOT. Ideal observer models of subcategorical information maintenance predict additive effects of phonetic cues and context on the log-odds of listeners' responses (Bicknell et al., 2025).

Power Analyses

Contrary to best practices, we did not conduct power analyses prior to experiments. Four previous experiments with a paradigm very similar to our high-informativity condition—for which we had conducted power analyses—suggested that our experiment was high powered to detect the effects of VOT and context after high-informativity exposure (Bicknell et al., 2025). This reasoning failed to take into account that the test phase in the present experiment is substantially shorter than in our previous experiments (48 vs. 168 trials) and that our previous experiments do not shed light on the expected effect size for the Context × Group interaction.

We thus conducted post hoc power analyses by repeatedly (1,000 times) generating data from the mixed-effects logistic regressions fit to our data. Data generation took into account all sources of variability present in the regression, including all random effects. Each data set we generated contained the same number of trials and participants as our actual data set. We then analyzed each of the 1,000 data sets with the same mixed-effects regression described above and calculated the proportion of times we found significant effects of context, VOT, and a significant Context × Exposure Group interaction at the $p < .05$ level. Two types of power analyses were conducted. The first analysis employed estimates based on previous work. Specifically, we set the effect of VOT to 6 log-odds per Gelman-scaled VOT unit (in four previous experiments with very similar paradigms, we found effects between 4.7 and 8.4; Bicknell et al., 2025; Bushong & Jaeger, 2019b) and the effect of context for high-informativity exposure to 1.6 log-odds (our previous experiments found effects between 1.02 and 2.08 log-odds). We note that this is a rather conservative setting as the mean of the context effect observed across our previous experiments—all of which employed informative exposure—is closer to 2 log-odds.

Since we had not previously conducted experiments on low-informativity exposure, we asked whether a halving in the context effect could be detected—that is, a context effect of 0.8 log-odds for low-informativity exposure. This, too, turns out to be conservative compared to the effects we observed in Experiment 1. All other effects were set to the (maximum-likelihood) estimates observed in Experiment 1. The second post hoc power analysis instead set all effects—including VOT, context, and the Context × Group interaction—to the values observed in Experiment 1. Table 1 reports the results of both power analyses.

Results

Figure 3 shows the average proportion of /t/-responses during the exposure and test phases by participant group, VOT, and subsequent context. Figure 4 shows the regression-estimated participant-specific effects of VOT and subsequent context in the exposure and test phases.

Exposure Phase

In line with Figure 3 (top left panel), the mixed-effects logistic regression found significant effects of both VOT ($\hat{\beta} = 6$, $z = 6.01$, $p < .001$) and subsequent context ($\hat{\beta} = 4.5$, $z = 6.68$, $p < .001$) on the first trial of the exposure phase: Participants were more likely to provide /t/-responses for larger VOTs and for *tent*-biasing, compared to *dent*-biasing, contexts. There was a main effect of trial, such that /t/-responses converged to 0 (50% in proportion space) over the course of the exposure phase ($\hat{\beta} = -0.15$, $z = -3.02$, $p < .01$). There was also a significant Trial × Context interaction, such that the context effect decreased over trials ($\hat{\beta} = -0.29$, $z = -2.96$, $p < .01$). The interaction between context and VOT was not significant ($p > .5$).

In order to estimate participant-specific context and VOT effects, we extracted the random slope estimates and standard errors for context and VOT for each participant. Participants were classified as having a significant effect if the 95% normal confidence intervals did not overlap 0. Figure 4 (leftmost panel) shows that the effects of VOT and subsequent context were present in the clear majority

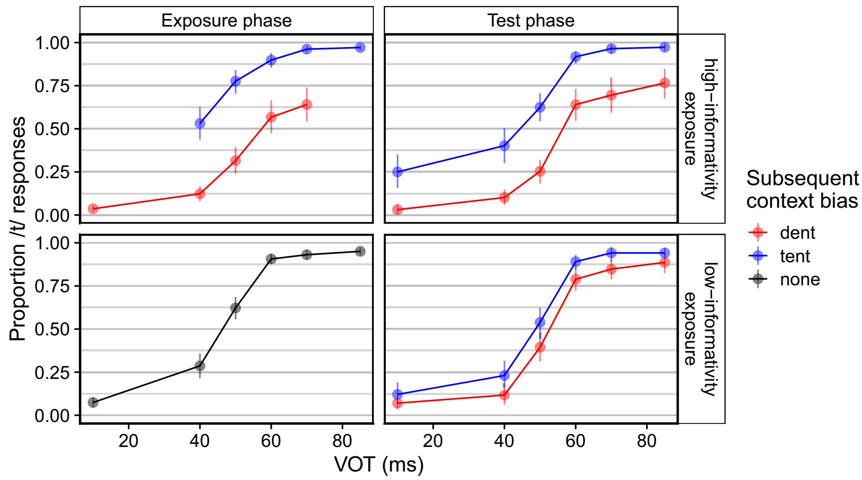
Table 1
Summary of Post Hoc Power Analyses for Experiment 1

Effect sizes based on	Previous work + halving of context effect for low-informativity exposure	Experiment 1
VOT	>99.9% (<.1%)	>99.9% (<.1%)
Context	90.7% (<.1%)	>99.9% (<.1%)
Context × Exposure Group	50.2% (<.1%)	96.4% (<.1%)

Note. Shown is the percentage of simulations that resulted in effects in the *predicted* direction (opposite of predicted in parentheses). Two types of analyses were conducted. See text for details. VOT = voice onset time.

Figure 3

Summary of Categorization Responses During the Exposure and Test Phase of Experiment 1 by VOT, Subsequent Context, and Exposure Group



Note. Error bars are bootstrapped 95% confidence intervals over by-participant means. VOT = voice onset time. See the online article for the color version of this figure.

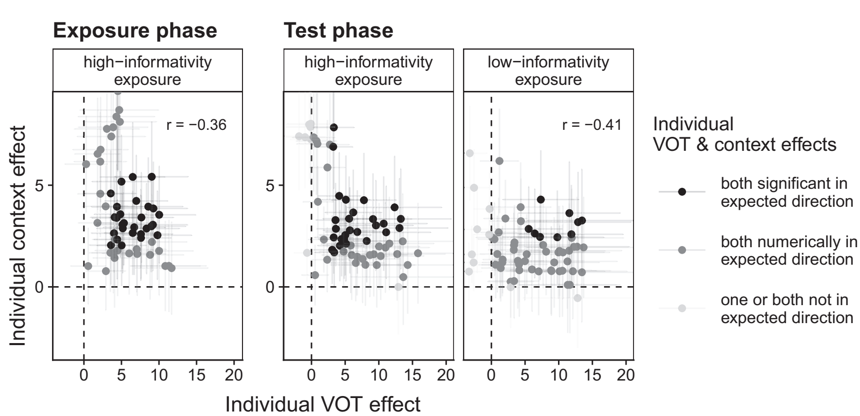
of participants. For the high-informativity condition, all of the 64 participants exhibited effects of both VOT and context in the expected direction, and these effects were "significant" even within participants for 28 (44%). We additionally conducted a simple Pearson correlation between participant-specific VOT and context effects; in the exposure phase, the correlation was negative ($r = -.36$).

Test Phase

We again found significant effects of VOT ($\hat{\beta} = 6.78$, $z = 9.02$, $p < .001$) and subsequent context ($\hat{\beta} = 2.78$, $z = 7.03$, $p < .001$) on the first trial of the test phase. As predicted, there was a significant interaction between context and group ($\hat{\beta} = -1.61$, $z = -3.49$, $p < .001$), such that the context effect was larger in the high-informativity group compared to the low-informativity group on the first trial of the test phase. Additionally, trial interacted with both context ($\hat{\beta} = -0.26$, $z = -3.53$, $p < .001$) and VOT ($\hat{\beta} = 0.41$, $z = 3.8$, $p < .001$), such that the context effect decreased and the VOT effect increased over the course of the test phase. No other effects or interactions were statistically significant, though there was a marginal interaction between group and VOT ($\hat{\beta} = 2$, $z = 1.77$, $p = .078$). This marginal effect goes in the opposite direction expected under any of the

Figure 4

Model-Predicted Participant-Specific Effects of VOT and Subsequent Context (in Log-Odds of /t/ Responses) Over the Exposure and Test Phase of Experiment 1



Note. Error bars show 95% confidence intervals. Participants who did not exhibit significant effects of both VOT and context (0 within 95% confidence intervals) are shown in lighter shades. Pearson's correlations of participant-specific random effects are shown (for both exposure and test) in the top right. VOT = voice onset time.

hypotheses outlined in the introduction (*larger* VOT effects after exposure to high-informativity contexts) and does not replicate in the other two experiments.

Paralleling the exposure phase, the majority of participants exhibited effects of both VOT and subsequent context (Figure 4, right panels). The percentage of participants who numerically exhibited both effects was similar in the two groups, but more participants in the high-informativity group exhibited significant effects of both VOT and context (low-informativity group: 84%, 14% significant; high-informativity group: 86%, 42% significant). As in the exposure phase, the VOT and context random effects were negatively correlated ($r = -.41$).

Discussion

With regard to our secondary research question, the results of Experiment 1 support the hypothesis that subcategorical information maintenance is the norm, rather than the exception. Participants in the high-informativity group showed effects of both VOT and context from the very beginning of the exposure phase, replicating our previous work (Bushong & Jaeger, 2019a). This was the case both across participants and within every single participant (Figure 4, left side). Indeed, compared to our previous experiments (Bicknell et al., 2025; Bushong & Jaeger, 2019a), which estimated the effects of context across all trials, the context effect estimated at the onset of the exposure phase in Experiment 1 is at least 50% larger. This result is expected if maintenance of subcategorical information is the *typical* behavior, rather than a reaction to the task demands of the experiment. This conclusion is also supported by the decrease in the effect of context over the course of the experiment: If anything, exposure to this type of experiment *reduces* the effects of subsequent context, rather than eliciting it.

With regard to our primary research question, the effect of subsequent context on listeners' responses during test was smaller after low-informativity exposure, compared to high-informativity exposure. At the same time, the effects of VOT were *not* significantly modulated by exposure: Listeners in both groups based their categorization responses to similar degrees on the VOT information in the acoustic signal. Both the reduced effect of subsequent context and the continued effect of VOT are predicted if a reduction in the expected informativity of subsequent context leads listeners to maintain subcategorical information less often—instead, categorizing the target word primarily (or solely) based on its VOT.

Interestingly, Figure 4 suggests that context effects were reduced—but still positive—in most participants. Such a pattern could arise, for example, if low-informativity exposure causes participants to use subsequent context less often (rather than never). The pattern would be less expected, for example, if low-informativity exposure only causes some random participants to be less likely to use subsequent context at all. We postpone further discussion of participant-specific patterns until the general discussion (the Maintenance of Subcategorical Information section).

Two potential concerns about Experiment 1 motivate Experiment 2. Both of these concerns relate to the fact that Experiment 1—like most previous work on subcategorical information maintenance—involved a large degree of repetition (see Bicknell et al., 2025; Bushong & Jaeger, 2017; Connine et al., 1991; Szostak & Pitt, 2013). This raises questions about the extent to which the results of Experiment 1 generalize to scenarios that more closely resemble the task demands of everyday language processing. In particular, participants in Experiment 1 were asked to make categorization judgments about the same critical target words of interest (*tent* and *dent*) throughout the entire experiment. This target word nearly always occurred as the third word of the sentence within the 12 different sentence contexts. It is possible that participants in Experiment 1 realized that it was sufficient to maintain subcategorical information about [t/d] (or about the relative probability of *tent* and *dent*), the only stimulus they were asked about. Experiment 1 might thus have overestimated participants' ability to maintain subcategorical information compared to everyday language use (see Burchill et al., 2018, for further discussion of this and related limitations). Experiment 2 takes a modest step toward addressing this issue by reducing the focus on the target word (as described in the next section). This change in the design also addresses a second concern about Experiment 1 that is of more immediate relevance to our present goal: For the high-informativity group, the knowledge that the critical target word was always *[t/d]tent* might have highlighted the fact that informative context always followed six to nine syllables after the target. It is possible that this inflated the differences in the context effect between the low- and high-informativity groups.

Experiment 2

As shown in Figure 1, Experiment 2 is identical to Experiment 1, except that participants in Experiment 2 only made judgments about the target words of interest (*tent* and *dent*) on half of the trials (randomly interspersed throughout the experiment). On the other half of trials, participants were instead asked to categorize another word in the sentence; these alternate target words were never the subsequent context words—that is, participants were never asked about the word "campgrounds" in Sentence 1a above (see the Supplemental Material for a list of all alternate target words). The goal of this change from Experiment 1 is twofold. First, it distributes participants' attention across more words in the stimulus sentences by making our word of interest *[t/d]ent* less salient in the task context. If listeners maintain subcategorical information only when they know exactly which segment or word this information needs to be maintained about, we should see a reduction in the context effect in Experiment 2. Second, the change in the design of Experiment 2 encourages participants in both exposure groups to remain attentive throughout the sentences, rather than simply "tuning out" after the target word. The change to Experiment 2 therefore should reduce potential differences between the low- and high-informativity groups that are due to which parts of the sentence listeners process, rather than which parts of the sentence listeners maintain subcategorical information about.

Method

Experiment 2 was approved by the University of Rochester Research Subjects Review Board. It was not preregistered.

Participants

As in Experiment 1, we recruited 128 native English-speaking participants for Experiments 2 (64 participants for each of the two between-participant conditions). Recruitment and reimbursement were identical to Experiment 1 and took place in November of 2018.

Participants who had participated in Experiment 1 were not allowed to participate in Experiment 2.

In Experiment 2, 43.7% of participants were female, 55.4% male, and 0.9% declined to respond. With respect to ethnicity, 9.4% reported as Hispanic, 89% as non-Hispanic, and 1.6% declined to respond. With respect to race, 87.5% of participants identified as White, 6.3% as Black or African American, 4.6% as Asian, 2.3% as American Indian/ Alaska Native or Native Hawaiian or other Pacific Islander, 1.5% as other, and 1.5% declined to respond. The mean age of our participants was 34.41 years (*SD* = 8.55, interquartile range = [29, 38]; 0.7% declined to respond).

Materials

The materials were identical to Experiment 1.

Procedure

The procedure was identical to Experiment 1 with one exception: On half of all exposure and test trials, participants instead had to categorize another word in the sentence (e.g., for Sentence 2c above, they were asked whether they heard "way" or "day").

Analysis Approach, Predictions, and Power

Analysis Approach

We analyzed Experiment 2 using the same methods as Experiment 1. The random effect structure of the exposure phase analysis was identical to Experiment 1. The random effect structure of the test analysis was identical to Experiment 1, except that the analysis of Experiment 2 only converged after the by-participant slopes for trial, and the by-item slopes for group and trial were excluded from the model. This is plausibly due to the fact that Experiment 2 had half the number of (critical) trials of Experiment 1.

Predictions

We again predicted significant effects of context and VOT, as well as the critical interaction between context and group, such that the effect of subsequent context is larger in the high-informativity group compared to the low-informativity group.

Power Analyses

Post hoc power simulations followed the exact same approach as in Experiment 1. Specifically, we repeated the same two power analyses as in Experiment 1 while setting the number of trials and participants, as well as all effect sizes other than context, VOT, and the Context × Group interaction to those observed in Experiment 2. Finally, we conducted a third analysis that set all effects—including those for context, VOT, and the Context × Group interaction—to those observed in Experiment 2.

As would be expected given our halving of the number of critical trials, power was reduced in Experiment 2. This was apparent both for the effect of context and for its interaction with exposure group. The loss in power was, however, small. This is due to the fact that participants, not items, were the primary source of variability: Variability in, for example, the context effect was an order of magnitude larger across participants (Experiment 1: $\hat{\sigma} = 2.31$; Experiment 2: $\hat{\sigma} = 2$) than across items (Experiment 1: $\hat{\sigma} = .23$; Experiment 2: $\hat{\sigma} = .16$, as estimated in the mixed-effects logistic regression analyses). Table 2 reports the results of the three power analyses.

Results

Figure 5 shows the average proportion of /t/-responses during the exposure and test phases by participant group, VOT, and subsequent context. Figure 6 shows the regression-estimated participant-specific effects of VOT and subsequent context in the exposure and test phases.

Exposure Phase

The results replicated Experiment 1: In line with Figure 5 (left), we found effects of VOT ($\hat{\beta} = 8.64$, $z = 5.16$, $p < .001$) and subsequent context ($\hat{\beta} = 3.75$, $z = 5.16$, $p < .001$) on the first trial of the experiment. Like Experiment 1, there was a significant negative main effect of trial ($\hat{\beta} = -0.16$, $z = -2.35$, $p = .01$). There was a marginal interaction between context and trial ($\hat{\beta} = -0.23$, $z = -1.73$, $p = .08$). Replicating Experiment 1, all participants numerically exhibited effects of both VOT and subsequent context, with 73% being significant (Figure 6, leftmost panel). The VOT and context random effects were negatively correlated ($r = -.3$).

Test Phase

The results of the test phase also largely replicated Experiment 1. We again found significant effects of VOT ($\hat{\beta} = 4.94$, $z = 7.67$, $p <$ .001) and subsequent context ($\hat{\beta} = 2.3$, $z = 4.82$, $p < .001$) on the first trial of the test phase. The interaction between subsequent context and exposure group was marginally significant and about half the size of the effect observed in Experiment 1 ($\hat{\beta} = -0.81$,

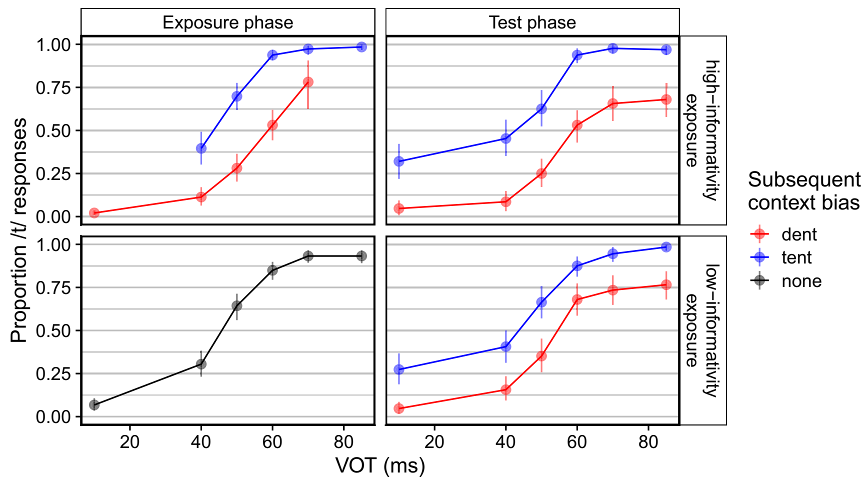
Table 2
Summary of Post Hoc Power Analyses for Experiment 2

Effect sizes based on	Previous work + halving of context effect for low-informativity exposure	Experiment 1	Experiment 2
VOT	>99.9% (<0.1%)	>99.9% (<.1%)	>99.9% (<0.1%)
Context	70.7% (<0.1%)	99.8% (<.1%)	>99.9% (<0.1%)
Context × Exposure Group	49.1% (5.7%)	94.6% (<.1%)	47.6% (3.6%)

Note. Shown is the percentage of simulations that resulted in effects in the *predicted* direction (opposite of predicted in parentheses). Three types of analyses were conducted. See text for details. VOT = voice onset time.

Figure 5

Summary of Categorization Responses During the Exposure and Test Phase of Experiment 2 by VOT, Subsequent Context, and Exposure Group



Note. Error bars are bootstrapped 95% confidence intervals over by-participant means. VOT = voice onset time. See the online article for the color version of this figure.

$z = -1.87$, $p = .06$). There were no other significant effects, including no interaction between exposure group and VOT.

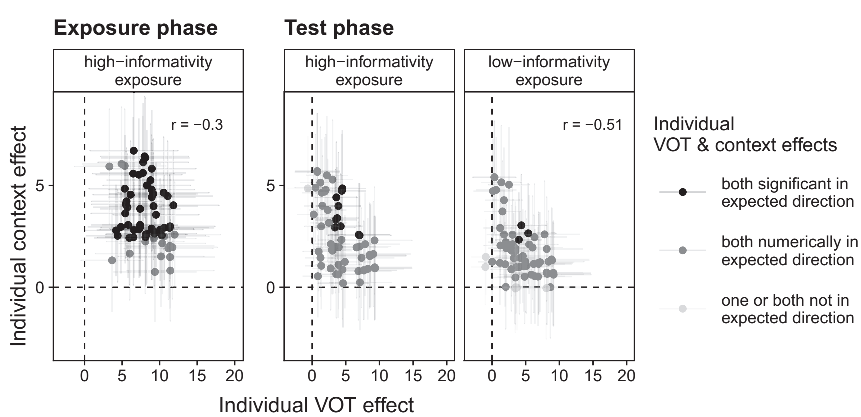
Also replicating Experiment 1 and paralleling the exposure phase of Experiment 2, we find that the majority of participants exhibited effects of both VOT and subsequent context in the numerically predicted direction (Figure 6, right). Echoing Experiment 1, the percentage of participants exhibiting both numerical effects was similar between groups, but more participants exhibited significant effects in the high-informativity group (high-informativity group: 98%, 17% significant; low-informativity group: 88%, 5% significant). As in the exposure phase, the VOT and context random effects were negatively correlated ($r = -.51$).

Discussion

Like Experiment 1, Experiment 2 found significant effects of context and VOT, suggesting that participants maintain subcategorical information about the target word beyond the word boundary. Unlike in Experiment 1, the Context × Group interaction was only marginally significant. A comparison of Figure 5 (Experiment 2)

Figure 6

Model-Predicted Participant-Specific Effects of VOT and Subsequent Context (in Log-Odds of /t/ Responses) Over the Exposure and Test Phase of Experiment 2



Note. Error bars show 95% confidence intervals. Participants who did not exhibit significant effects of both VOT and context (0 within 95% confidence intervals) are shown in lighter shades. Pearson's correlations of participant-specific random effects are shown in the top right. VOT = voice onset time.

against Figure 3 (Experiment 1) also suggests that the *size* of some of these effects differed across the two experiments. We thus conducted a post hoc analysis over the combined data from both experiments. This analysis—reported in full in the Supplemental Material (Section 3)—was identical to the mixed-effects logistic regression analyses for Experiments 1 and 2 but additionally includes experiment (sum-coded) and all of its interactions. This post hoc analysis found that the VOT effect in Experiment 2 was significantly smaller than in Experiment 1 ($p < .001$). There was a significant three-way interaction between experiment, context, and trial ($p = .02$), such that the context effect was larger in Experiment 2, compared to Experiment 1, as the test phase progressed. The three-way interaction between exposure group, context, and experiment was not significant: While the difference in the context effect between the high- and low-informativity groups was smaller in Experiment 2 ($\hat{\beta} = -0.81$ compared to $\hat{\beta} = -1.61$ in Experiment 1), it was not significantly smaller ($p = .41$).

Two of these results deserve discussion. First, the VOT effect was significantly smaller in Experiment 2. Second, the Context × Group interaction was numerically—but not significantly—smaller in Experiment 2. The reduced effect of VOT suggests that participants in Experiment 2 relied more on context and less on VOT, compared to Experiment 1. One intuitive explanation for this is that the change in the design of Experiment 2 caused participants to pay more attention to the subsequent context, causing them to pay less attention to the VOT. Critically, even in Experiment 2, the vast majority of all participants exhibited effects of *both* VOT and context—that is, the combination of effects that indicates that participants maintained subcategorical information, rather than to just respond based on either the VOT or the subsequent context (Figure 6). Indeed, in Experiment 2, all but six participants exhibited effects of VOT and context in the predicted direction. These subtle differences between experiments thus seem to point to quantitative differences in the use of VOT and context by participants, not qualitative differences in subcategorical maintenance behavior.

The second difference between Experiments 1 and 2 is that the Context × Group interaction was only marginally significant in Experiment 2. The combined analysis showed that the interaction was *not* significantly smaller in Experiment 2. On the one hand, a null effect is not convincing evidence for the absence of differences between the experiments. On the other hand, there is a good reason to expect the difference in significance between experiments: Since participants in Experiment 2 only provided responses of interest on half of all trials, Experiment 2 is based on half the amount of data as in Experiment 1. It is plausible that reduced statistical power, rather than underlying differences in the effects of context, explains why the effect in Experiment 2 was only marginally significant.

There are additional design concerns about both Experiment 1 and 2 that we raised in the description of Experiment 1. First, in half of the experimental lists, participants heard the same items in exposure and test. For the high-informativity condition, this meant that those participants heard the exact same sentences, whereas this was never the case for the low-informativity condition (since exposure employed the neutral and test employed the *dent*- or *tent*-biasing variants of the item). It is possible that half of the participants who saw repeated sentences inflated the context effect for the high-informativity group during the test phase, causing the observed difference in context effects between the exposure conditions. To assess this possibility, post hoc analyses reported in the Supplemental Material (Section 4) compared the context effects for participants in the high-informativity condition depending on whether or not they saw exposure sentences repeated during test. While context effects for repeated items were minimally smaller, this interaction was not statistically significant (Experiment 1: $\hat{\beta} = -0.34$, $z = -0.55$, $p = .58$; Experiment 2: $\hat{\beta} = -0.08$, $z = -0.15$, $p = .88$).

Finally, a potentially more problematic concern is the difference in the number of sentence repetitions during exposure between the high- and low-informativity conditions. Recall that items for Experiments 1 and 2 consisted of a triplet of sentences: a *tent*-biasing context variant, a *dent*-biasing context variant, and a neutral-context variant. While our design repeated exposure *items* equally often in both conditions, the number of repetitions for each *sentence* was twice as large in the low-informativity group. It is possible that this caused participants in the low-informativity group to disengage as the experiment advanced, leading to reduced attention and reduced context effects during the test phase. To address these potential confounds and concerns about statistical power, we conducted Experiment 3.

Experiment 3

Experiment 3 is a conceptual replication of Experiment 2 with double the number of recruited participants. We additionally made two design changes motivated by the concerns discussed above.

Method

Experiment 3 was approved by the University of Rochester Research Subjects Review Board. It was not preregistered.

Participants

We recruited 256 native English-speaking participants (128 in each between-participant condition). Participants were recruited in September 2022 from Amazon Mechanical Turk and rewarded $5.00 for their participation (for a targeted rate of $10.00/hr). Against instructions, nine participants took the experiment multiple times. We kept the first instance of each participant but removed their extra data, leaving data from 237 unique participants for analysis (92.6%).

We collected demographic information per the National Institutes of Health reporting requirements. In Experiment 3, 44.7% of participants were female, 54.9% male, and 0.4% declined to respond. With respect to ethnicity, 6.3% reported as Hispanic, 91.1% as non-Hispanic, and 2.6% declined to respond. With respect to race, 90.2% of participants identified as White, 4.6% as Black or African American, 3.8% as Asian, 1.2% as American Indian/Alaska Native or Native Hawaiian or other Pacific Islander, .4% as other, and .8% declined to respond. The mean age of our participants was 37.9 years ($SD = 10.9$, interquartile range = [30, 44]; 0 declined to respond).

Materials

In order to reduce repetition of sentences in the low-informativity condition, we utilized additional neutral context sentences beyond those used in Experiments 1–2. We had eight additional neutral context sentences that were recorded from the same talker in the same session as the materials for Experiments 1 and 2. These new items were added to the pool of items to be used in the low-informativity context exposure group (see the Appendix for the full list of stimuli and items used in Experiments 1–3).

Procedure

The procedure was identical to Experiment 2 with the exception of the construction of experimental lists (opaque to participants). For Experiment 3, we made two minor modifications to the structure of items and trials. First, Experiment 3 never repeated items between exposure and test.

Second, participants in the low-informativity condition saw the same number of unique exposure sentences—and experienced the same degree of repetition—as participants in the high-informativity condition. In Experiments 1 and 2, we used six sentence items in the exposure phase. This meant that subjects in the high-informativity condition heard 12 unique sentences (6 items × 2 context bias conditions), each repeated six times, but subjects in the low-informativity condition only heard six unique sentences (6 items × 1 neutral context), each repeated 12 times. In Experiment 3, participants in the low-informativity condition instead heard 12 unique sentences (12 items × 1 neutral context), each repeated six times. As in Experiments 1 and 2, we constructed several lists to counterbalance which items appeared in exposure versus test.

Analysis Approach, Predictions, and Power

Analysis Approach

We analyzed Experiment 3 using the same approach as Experiments 1 and 2. The random effect structure was identical to Experiment 2.

Predictions

As in Experiments 1–2, we expected to find significant effects of context and VOT, as well as an interaction between context and group, such that the effect of subsequent context is larger in the high-informativity group compared to the low-informativity group.

Power Analyses

Post hoc power simulations followed the same approach as in Experiments 1 and 2. Specifically, we repeated the same three power analyses as in Experiment 2—effect estimates based on previous work, effect estimates based on Experiment 1, and effect estimates based on Experiment 2—while setting the number of trials and participants, as well as all effect sizes other than context, VOT, and the Context × Group interaction to those observed in Experiment 3. As would be expected given that we doubled the number of subjects compared to Experiment 2, power was estimated to be higher for all effects of interest in Experiment 3 compared to Experiment 2. Estimated power for all predicted effects was ≥80%. Table 3 reports the results of the three power analyses.

Results

Figure 7 shows the average proportion of /t/-responses during the exposure and test phases by participant group, VOT, and subsequent context. Figure 8 shows the regression-estimated participant-specific effects of VOT and subsequent context in the exposure and test phases. There are two immediate observations based on these figures. On the one hand, the results of Experiment 3 seem to follow the same qualitative pattern as for Experiments 1 and 2, with a positive effect of VOT, and a main effect of context that is reduced after low-informativity exposure. On the other hand, Figure 7 reveals striking differences compared to Figures 3 and 5 from Experiments 1 and 2: Both the context effect and the effect of VOT seem to be reduced in Experiment 3. As we show below, both of these differences were caused by inattentive participants in Experiment 3. Before we address this issue, we present our planned analyses. These analyses follow the exact same approach as in Experiments 1 and 2 and thus do not introduce additional exclusion criteria. Following those analyses, we confirm that all results of Experiment 3 replicate—and indeed become stronger—when inattentive participants are excluded. This makes sure that our results hold *regardless of the exclusion criteria employed*.

Exposure Phase

Conceptually replicating Experiments 1 and 2, the planned analyses found main effects of VOT ($\hat{\beta} = 1.21$, $z = 2.25$, $p = .02$) and subsequent context ($\hat{\beta} = 2.11$, $z = 4.34$, $p < .001$) on the first trial of the experiment. No other main effects or interactions were significant. In line with Figure 7, the estimated coefficients of both VOT and context were substantially smaller than in Experiments 1 (VOT: $\hat{\beta} = 6.00$; context: $\hat{\beta} = 4.5$) and 2 (VOT: $\hat{\beta} = 8.64$; context: $\hat{\beta} = 3.75$). Also conceptually replicating Experiments 1 and 2, the majority of participants exhibited significant effects of both VOT and subsequent context, though the proportions were notably lower (69% numerically predicted direction, 12% significant; Figure 8, leftmost panel).[6] Unlike in Experiments 1 and 2, the random effects of VOT and context were positively correlated ($r = .5$).

Test Phase

The results of the test phase also conceptually replicated all critical effects of Experiment 1, though again with smaller coefficient estimates. We again found significant effects of VOT ($\hat{\beta} = 1.89$, $z = 3.84$, $p < .001$) and subsequent context ($\hat{\beta} = 0.93$, $z = 3.14$, $p = .002$) on the first trial of the test phase. As in Experiment 1, we observed a significant interaction between context and group ($\hat{\beta} = -0.55$, $z = -2.18$, $p = .03$), such that the context effect was lower in the low-informativity group. Additionally, there was a marginal main effect of group ($\hat{\beta} = 0.52$, $z = 1.69$, $p = .09$) such that the high-informativity group was more likely to respond /t/, and a significant interaction between group and trial ($\hat{\beta} = -0.13$, $z = 3.14$, $p = .02$). There were no other significant effects, including no interaction between exposure group and VOT.

Also conceptually replicating Experiments 1 and 2, we find that the majority of participants exhibited effects of both VOT and subsequent context in the numerically predicted direction (Figure 8, right). Notably, however, the proportions were much smaller in both groups (high-informativity group: 64%, 2% significant; low-informativity

[6] The low percentage of participants with significant effects of both VOT and subsequent context is in part a consequence of the relatively large proportion of inattentive participants in Experiment 3 (i.e., the cluster of points around the origin in Figure 8). Not only do these participants not exhibit VOT or context effects, they also pull the estimated effects of all *other* participants toward 0 (due to the "shrinkage" effects of mixed-effects regressions, as confirmed in Supplemental Material, Section 5).

Table 3
Summary of Power Analyses for Experiment 3

Effect sizes based on	Previous work + halving of context effect for low-informativity exposure	Experiment 1	Experiment 2
VOT	>99.9% (<.1%)	>99.9% (<.1%)	>99.9% (<.1%)
Context	91.5% (<.1%)	>99.9% (<.1%)	>99.9% (<.1%)
Context × Exposure Group	80% (<.1%)	>99.9% (<.1%)	81.6% (<.1%)

Note. Shown is the percentage of simulations that resulted in effects in the *predicted* direction (opposite of predicted in parentheses). Three types of analyses were conducted. See text for details. VOT = voice onset time.

group: 59%, 2% significant). As in the exposure phase, the random effects of VOT and context were positively correlated, though more weakly ($r = .14$).

Discussion

The planned analyses for Experiment 3 replicate all effects of Experiment 1 for a design that is a conceptual replication of Experiment 2. However, while significant in the expected directions, all critical effects in Experiment 3 were of smaller magnitude than in Experiment 1. Given reports of increasing proportions of inattentive or uncooperative participants on the Mechanical Turk platform (Peer et al., 2017, see also Supplemental Material, Section 5.1), we suspected that some participants in Experiment 3 might have responded randomly. This would straightforwardly explain the reduced effects as a mixture of the true effects and the null effects of random responders.

Indeed, there is clear evidence that Experiment 3, which was conducted more than 3 years after Experiments 1 and 2, recruited a less attentive pool of participants. Both Experiment 2 and Experiment 3 provide us with a simple measure of participants' attentiveness: the accuracy that participants achieved on noncritical/filler trials. On those trials, participants were asked whether they heard, for example, "way" or "day" in example filler (2c) above. The stimulus for the correct answer ("way" in Example 2c) was never phonetically manipulated and always easy to understand. Figure 9 shows a dramatic drop in participants' accuracy on those filler trials in Experiment 3, compared to Experiment 2 (no comparable data were collected in Experiment 1). In particular, a nontrivial subset of participants in Experiment 3 performed around chance level (50%).[7]

Given this pattern, we decided to conduct a post hoc robustness analysis to ensure that our results hold when inattentive participants are excluded. Since we had not anticipated this issue and had thus not committed to planned threshold values for attentiveness, we tested *all possible* thresholds. The goal of this approach is to ensure that the results of our planned analyses hold under *any* exclusion threshold for participants' accuracy on filler trials. This also allows us to test a critical prediction: If the reduced coefficient estimates in Experiment 3 are indeed caused by inattentive participants, coefficient estimates for predicted effects should steadily increase with more conservative thresholds for attentiveness (at least up to a point at which too few participants are available for analysis to obtain reliable estimates).

Figure 10, which summarizes the results of this exclusion analysis (for details, see Supplemental Material, Section 5), suggests that this prediction is confirmed: More attentive participants also exhibited larger effects of VOT, subsequent context, and the critical interaction between exposure condition and context. Specifically, the coefficient estimates for all three effects converged toward those observed in Experiment 1—a systematic tendency that would be highly unexpected by chance. The Supplemental Material (Section 5) provides additional versions of Figure 7 at different exclusion thresholds, which further illustrate this point. The same section of the Supplemental Material also shows that the clear majority of attentive participants exhibited effects of both VOT and context, as well as a negative correlation between the two effects, replicating Experiments 1 and 2. This suggests that differences between Experiments 1–2 and 3 can be reduced to the presence of inattentive participants.

Taken together, our planned and post hoc analyses suggest that the differences in statistical significance between Experiments 1 and 2 were primarily due to differences in statistical power, rather than actual differences in results—in line with the comparison of the two experiments we presented in the discussion of Experiment 2. Finally, Experiment 3 addressed two potential concerns about Experiments 1 and 2 that might be seen as confounding the effect of informativity. First, none of the test sentences were repetitions of exposure sentences in Experiment 3. Second, the degree of sentence repetition during exposure was held constant across the two exposure conditions. Experiment 3 replicates the significantly reduced effect of subsequent context after low-informativity exposure in the absence of these confounds.

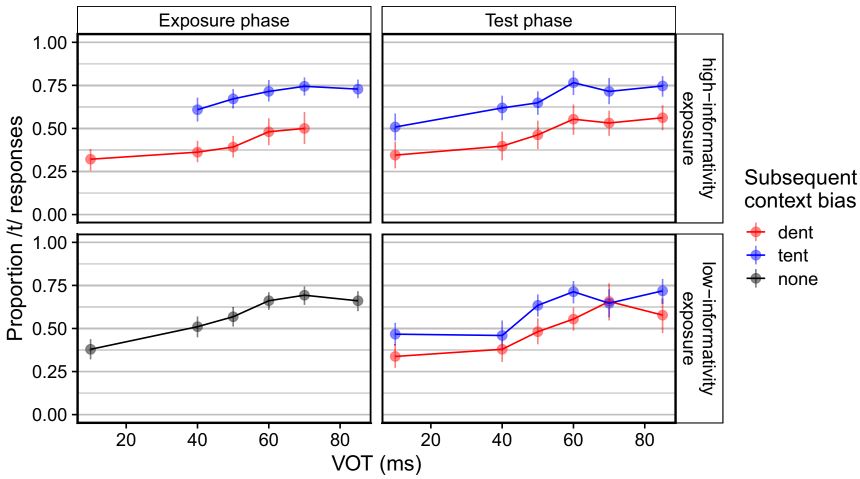
General Discussion

Listeners process thousands of bits of incoming information per second of speech, compressing the speech signal into linguistic representations and, ultimately, meanings. However, subcategorical information about preceding speech input can be useful to maintain in memory for integration with subsequent input. The present study adds to work that has found listeners can maintain subcategorical representations of previous input beyond the word boundary (Brown-Schmidt & Toscano, 2017; Burchill et al., 2018; Bushong & Jaeger, 2019a; Connine et al., 1991; Falandays et al., 2020; Szostak & Pitt, 2013). This maintenance process is not unlimited, however, raising the question of how listeners decide what subcategorical information to maintain and when (Burchill, 2023; Burchill et al., 2018; Caplan et al., 2021; Christiansen & Chater, 2016). We proposed that listeners might adapt the degree to which they maintain subcategorical information based on the expected informativity of subsequent context. We tested this by manipulating the informativity of subsequent context: If subsequent context is likely to be informative for accurate recognition, then the expected utility of maintaining subcategorical

[7] The proportion of inattentive participants did not differ across conditions of Experiment 3 (as confirmed by a mixed-effects logistic regression predicting accuracy on filler trials from exposure condition, $p > .98$).

Figure 7

Summary of Categorization Responses During the Exposure and Test Phase of Experiment 3 by VOT, Subsequent Context, and Exposure Group



Note. Error bars are bootstrapped 95% confidence intervals over by-participant means. VOT = voice onset time. See the online article for the color version of this figure.

information about VOT is higher, since it will be available for integration with subsequent context.

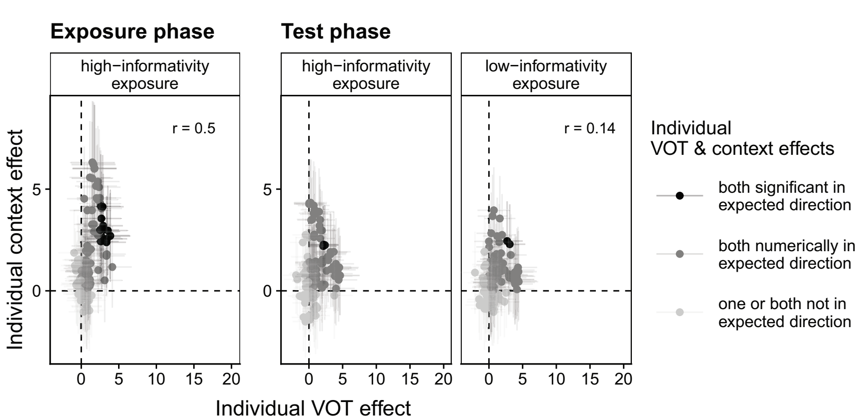
We begin by discussing what our results suggest about the maintenance of subcategorical information. As part of this, we revisit the assumptions made both in our and in previous research on subcategorical information maintenance. Following this, we discuss the results of the informativity manipulation.

Maintenance of Subcategorical Information

Studies on the maintenance of subcategorical information—including our own work—have typically proceeded under the assumption that it is sufficient to demonstrate effects of both the subcategorical information *and* the subsequent context. When both of these effects are observed, this is taken to constitute evidence for subcategorical information maintenance (e.g., Bicknell et al., 2025; Brown-Schmidt & Toscano, 2017; Bushong & Jaeger, 2017; Connine et al., 1991; Falandays et al., 2020; Szostak & Pitt, 2013). Take, for example, the beginning of the sentence "After the ?ent Sue had found in the campgrounds collapsed, ..." A listener who does not maintain subcategorical information is thought to categorize the "?" sound into /d/ or /t/ based on the information available at that moment (e.g., the VOT) and then to immediately discard all subcategorical information which led them to that decision. Later, when the listener encounters "campgrounds," they retain their earlier decision because they have no access to subcategorical

Figure 8

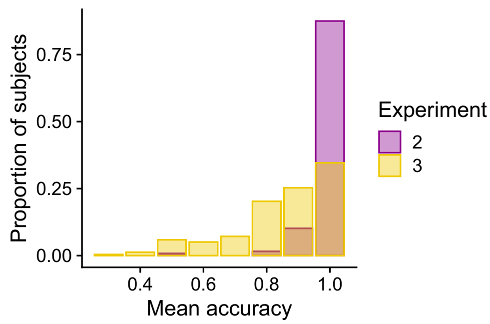
Model-Predicted Participant-Specific Effects of VOT and Subsequent Context (in Log-Odds of /t/ Responses) Over the Exposure and Test Phase of Experiment 3



Note. Error bars show 95% confidence intervals. Participants who did not exhibit significant effects of both VOT and context (0 within 95% confidence intervals) are shown in lighter shades. Pearson's correlations of participant-specific random effects are shown in the top right. VOT = voice onset time.

Figure 9

Distribution of By-Participant Accuracy on Filler Trials in Experiments 2 and 3



Note. Filler trials did not contain manipulated phonetic cues and thus were always easy to recognize accurately. Note that Experiment 2 has approximately half the number of subjects as Experiment 3. See the online article for the color version of this figure.

information from the previous input with which to integrate this new input. Thus, at the end of the trial, categorization responses reflect the earlier subcategorical information only. If, instead, one observes a pattern where both subcategorical information and later context are used, this is taken to reflect integration of both cues, implying that subcategorical information has been maintained.

Under these assumptions, the results of Experiments 1–3 would seem to show that listeners maintain subcategorical information for at least six to nine syllables—the same interpretation that previous work reached for similar evidence. However, as we anticipated in the introduction, the presence of effects of both subcategorical and subsequent contextual information is not actually sufficient to conclude that listeners maintain subcategorical information. Here, we discuss two broad classes of alternative explanations and the extent to which existing evidence is compatible with them. The first type of explanation appeals to the idea that listeners can *revise* their categorization decisions and that this can explain the effects of both VOT and context. This idea has theoretical precedence in the literature on sentence processing (e.g., the two-stage model, Frazier & Fodor, 1978). The second type of explanation instead focuses on the idea that the effects of VOT and context originate in a *mixture* of strategies that rely on either only VOT or only subsequent context.[8]

Turning to the first of these alternative explanations, would the presence of VOT and context effects not also be explained if listeners revise their decisions based on later information, without retention of earlier input? Assume, for example, that on a particular trial, the listener categorizes a segment as /d/ based on the acoustic evidence and then discards all subcategorical information which led them to that decision. When they reach "campgrounds," this conflicts with their initial decision that they heard the word "dent" earlier. Listeners could then *switch* their categorization to /t/. In this way, listeners need not maintain any subcategorical information about previous input, but their behavioral responses would appear to reflect both early subcategorical information and later context. In previous work, we formalized this *categorize-discard-and-switch* hypothesis and implemented it as a model to derive detailed quantitative predictions (Bushong, 2020; Bushong & Jaeger, 2019c). We confirmed that this switching model indeed predicts effects of both VOT and context. The specific way these two effects combine does, however, differ from a model of ideal subcategorical information maintenance (developed and described in detail in Bicknell et al., 2025). Specifically, the categorize-discard-and-switch model predicts a sublinear interaction between VOT and context, whereas ideal information maintenance predicts additive effects of VOT and context (for details, see Bushong, 2020). When we compared both models against data from four experiments similar to the ones presented here, the switching model consistently provided a poor fit to listeners' behavior. If these results are replicated in future research, this would suggest that the categorize-discard-and-switch hypothesis is *not* the most likely explanation for results like the ones obtained here.[9]

The second type of alternative explanation comes in at least two variants. First, it would be possible that some participants always use only VOT to categorize our stimuli, and other participants always use context. Let us call this the *participant-mixture hypothesis.* When aggregating across participants, this scenario can result in effects of both VOT and subsequent context, even if none of the participants ever maintain subcategorical information. In previous work, we addressed this issue by excluding participants who did not show an effect of VOT from analysis (e.g., Bushong & Jaeger, 2017, 2019c). This ensures that any effect of subsequent context in those experiments occurred in the presence of effects of VOT. This approach does, however, make it impossible to estimate how pervasive or typical subcategorical information maintenance is across listeners. Here, we instead took an alternative approach—one that we feel is more transparent and more informative. We derived participant-specific estimates of both VOT and context effects from our mixed-effects analyses (see also Bicknell et al., 2025). This made it possible to assess whether maintenance of subcategorical information is the default both within and across participants, at least in the type of experiment conducted here.

This made apparent that some participants relied more on context, while others relied more on VOT, resulting in the negative correlations between the participant-specific effects of VOT and context in Experiments 1–3 (for Experiment 3, this pattern held only for attentive participants; see Supplemental Material, Section 5). Overall, the effects of VOT and context trade off against each other, as evidenced by the negative correlations between the two effects (see also Bicknell et al., 2025). Such trade-off relations are common for any type of cue integration and are known to vary across individuals (e.g., Kim & Clayards, 2019; Schertz & Clare, 2020). To the extent that these differences reflect true individual differences, they might reflect differences in participants' strategies to deal with the task demands of the paradigm (and/or boredom). Alternatively, this pattern might point to individual differences in the *ability* to maintain subcategorical information. Experiments with substantially more trials—and thus more data per participant—would be required to address this question convincingly.

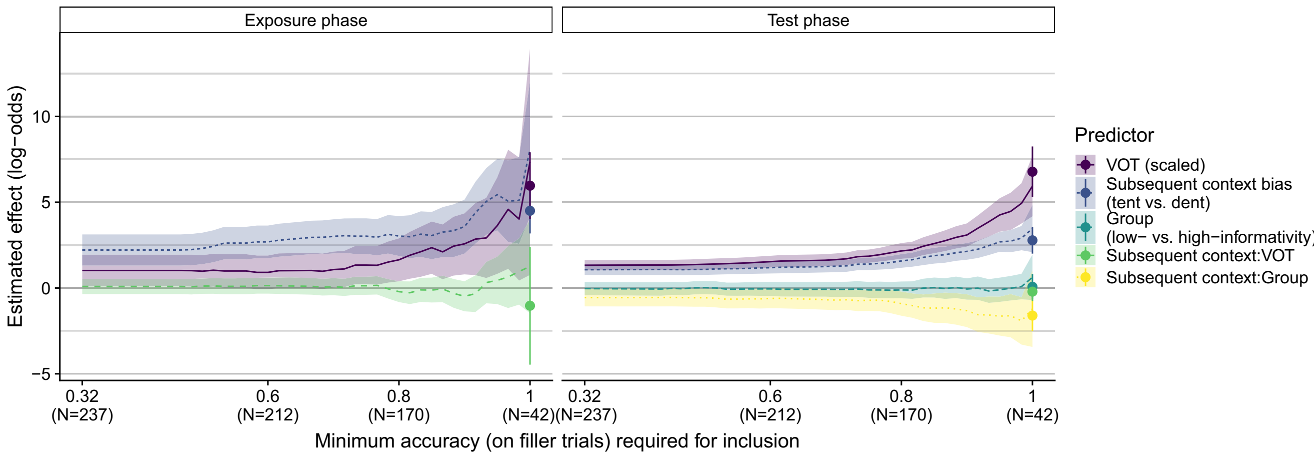
Critically, the vast majority of participants of Experiments 1 and 2 in both the high- and low-informativity exposure groups displayed effects of both VOT and context in the numerically predicted

[8] We are grateful to two anonymous reviewers for constructively engaging with this discussion and helping to advance it. All remaining oversights remain our own.

[9] In the introduction, we also raised the possibility that listeners might *reduce* the maintenance of subcategorical information by switching to the categorize-discard-and-switch strategy on some or all trials. This is a separate hypothesis addressed in the next section.

Figure 10

Coefficient Estimates With 95% Confidence Intervals (y-Axis) by Minimum Response Accuracy on Filler Trials (x-Axis)



Note. Estimates are derived from mixed-effects logistic regressions fitted to different subsets of participants, successively excluding participants who did not reach minimum filler accuracy (x-axis). The point ranges on the right-hand side of each panel indicate coefficient estimates with 95% confidence intervals from Experiment 1 for comparison. Note how coefficient estimates in Experiment 3 tended to converge toward the estimates obtained in Experiment 1, as more and more inattentive participants were excluded. VOT = voice onset time. See the online article for the color version of this figure.

direction (90% across both experiments' test phases). Despite the substantial reduction in statistical power inherent to analyzing participant-specific effects, the effects of VOT and context reached significance for a sizable proportion of individual participants (20%). This suggests that overall, our participants are capable of maintaining subcategorical information about VOT in memory for later integration with context. In Experiment 3, we found the same pattern for participants who attended to the task. Generally, it was thus *not* the case that some participants only used context and others only used VOT. Rather, most participants used both VOT and context, as expected under the hypothesis that they are maintaining subcategorical information.

However, even if most participants exhibit effects of both VOT and context, this still does not necessarily entail subcategorical information maintenance. The reasoning of the participant-mixture hypothesis can also be applied at the trial level: Any individual participant might rely on only VOT on some trials and rely on only context on other trials. This *trial-mixture hypothesis*, too, would yield effects of both VOT and context in the absence of subcategorical information maintenance.

Unlike for the categorize-discard-and-switch hypothesis and the participant-mixture hypothesis, existing data—including Experiments 1–3—would seem to be perfectly compatible with the trial-mixture hypothesis. Indeed, we anticipate that it will take nonnegligible experimental effort to address this hypothesis and distinguish it from the hypothesis that listeners maintain subcategorical information. Specifically, future research could use the same set of stimuli to (a) estimate the effects of VOT in the absence of subsequent context, (b) estimate the effects of subsequent context in the absence of VOT, and (c) compare them to the effects of VOT and subsequent context when the entire sentence stimulus is used (as in Experiments 1–3). Similar approaches have been employed in research on cue integration (e.g., Bejjanki et al., 2011). At least under the hypothesis of *ideal* subcategorical information maintenance—that is, if listeners maintain subcategorical information perfectly (without loss of information) and integrate it with subsequent context based on the relative reliability of each cue (Bicknell et al., 2025)—then the effects of VOT and context should be identical to the effects of each cue in the absence of the other cue.

Such research would have to carefully address several likely confounds. This includes, for example, concerns about coarticulation (acoustic evidence for /d/ vs. /t/ might well be available prior to the onset of the target word), as well as the possibility that cue conflicts between VOT and context in condition (c) can change the effects of either cue over the course of the experiment (Bushong & Jaeger, 2019a) in ways that would not be observed in conditions (a) and (b). Researchers also would have to address potential concerns that differences in the interpretation of the task for conditions (a)–(c), or resulting differences in the time course of responses, will drive differences in the effects of VOT and context. Still, we consider the comparison outlined here a promising avenue for future research.

In summary, under the assumptions made in previous research on this topic, the presence of both VOT and context effects in Experiments 1–3 constitutes evidence for subcategorical information maintenance. There are, however, alternative explanations for these effects that have not previously received much attention. We have discussed two types of alternative explanations. While one of these alternatives—the categorize-discard-and-switch hypothesis—seems to be disfavored by existing data, one variant of the other alternative—the trial-mixture hypothesis—seems to be compatible with existing data, explaining those data without reference to maintenance of subcategorical information. This is an important consideration for future research on subcategorical information maintenance and beyond.

Effect of Informativity on the Use of Context in Spoken Word Recognition

In this final discussion section, we set aside uncertainty about the causes for the effects of VOT and subsequent context and discuss what the effects of exposure observed in Experiments 1–3 would entail *under the standard assumptions made in previous research on subcategorical information maintenance*.

Recall that we found reduced effects of subsequent context on participants' responses in the low-informativity group compared to the high-informativity group. This reduction was significant in Experiments 1 and 3 and marginally significant in Experiment 2. However, the design of Experiment 1 shares with previous work (Bicknell et al., 2025; Bushong & Jaeger, 2019a; Connine et al., 1991; Szostak & Pitt, 2013) that it focused participants' attention on a single aspect of our sentence stimuli, the target word (or perhaps even just the onset sound of the target word). This left open whether Experiment 1 and previous work might be overestimating the effects of VOT and/or context (see discussion in Burchill et al., 2018). Experiment 2 was designed to begin to address this question. Participants did not always make judgments about our critical target words of interest. Instead, on half of all trials (during exposure and test), they made categorization judgments about noncritical words in the sentence. This change in the design takes a (small) step toward the task demands of natural language use: Listeners do not necessarily a priori know which parts of the speech input they will need to comprehend and/or respond to. When we made this change, we observed similar qualitative results to Experiment 1, though the interaction between context and group was only marginally significant. This raised the question whether the loss of significance in Experiment 2 was driven by reduced power, due to the fact that Experiment 2 had half as many critical trials as Experiment 1. Following reviewers' advice, we conducted Experiment 3 to replicate the (marginally significant) interaction in Experiment 2. After removing inattentive participants, we indeed found the same significant context by exposure group interaction in Experiment 3 as in Experiment 1, with reduced effects of subsequent context after low-informativity exposure.

This reduction of the context effect was the key prediction we made in the introduction under the hypothesis that listeners modulate the extent to which they maintain subcategorical information depending on its expected utility. In the low-informativity exposure condition, informative subsequent context is not available. We hypothesized that this would lead listeners to expect that there would be little utility to maintaining subcategorical information, so that listeners should increasingly start to respond based on only VOT. For Experiments 1–3, we would thus predict that participants in the low-informativity group would exhibit a reduced effect of subsequent context. Specifically, if listeners in the low-informativity group stop maintaining subcategorical information altogether, we would expect their context effect to be exactly zero. If listeners instead reduced, but did not completely stop maintaining, subcategorical information, we would expect reduced but nonzero effects of subsequent context. The latter is what we observed in Experiments 1–3.

In the introduction, we also anticipated two alternative ways in which listeners may reduce subcategorical information maintenance in Experiments 1–3. One possibility is that listeners in the low-informativity group might begin to adapt the categorize-discard-and-switch strategy. This account would predict reduced effects of *both* VOT and context in the low-informativity group. Alternatively, listeners might stop maintaining subcategorical information on some or all trials, relying instead on only VOT during those trials. Unlike the first possibility, this second possibility does not predict reduced effects of VOT. The results of Experiments 1–3 favor the latter of these two interpretations: None of the three experiments found any evidence that low-informativity exposure leads to reduced effects of VOT.

The absence of changes in the effect of VOT also argues against a third possibility that listeners reduce the maintenance of subcategorical information by allowing subcategorical information to decay more quickly in short-term memory. This could be understood as introducing more noise into listeners' short-term memory representations. This hypothesis would predict (a) a smaller effect of VOT in the low-informativity group (as noisier or otherwise decayed representations of VOT/uncertainty would reduce the informativity of VOT; Feldman et al., 2009) and (b) identical context effects for both exposure groups (since both groups of listeners continue to maintain and integrate subcategorical information with the subsequent context). For Experiments 1–3, this strategy would *not* appear to be a rational response to the low-informativity condition, since the informativity of VOT did not differ across exposure conditions. Indeed, neither (a) nor (b) were observed in Experiments 1–3. Rather, the present results point to a relatively simple mechanism for listeners in the low-informativity group, which is to stop maintaining subcategorical information altogether on at least some of the trials and rely on initial decisions based on early cues (here, VOT).

One final point deserves discussion. Experiments 1–3 manipulated the availability of context as an informative cue to word identity. This bears some similarity to previous work that has manipulated the *reliability* of cues in the environment (e.g., Bushong & Jaeger, 2019a; Idemaru & Holt, 2011). For example, in Bushong and Jaeger (2019a), we manipulated how often participants encountered context that conflicted with the earlier VOT cue in the sentence. The more that cues conflicted, the more listeners down-weighted context in their categorization responses. At first blush, this parallels the results we observe here: a manipulation of context and resultant reduction in use of context. However, the present results were obtained in the absence of any changes in the reliability of subsequent context. During exposure, the low-informativity group receives no evidence one way or the other that context may be unreliable with respect to category identity, just that context is usually uninformative about category identity. And, during the test phase, both groups of participants saw the exact same trials, in which subsequent context was informative (and just as reliably so as during high-informativity exposure). Theories of cue reweighting would thus not predict that the participants in the low-informativity group should change the way they use context during the test phase.

Conclusion

The present work suggests that listeners are sensitive to the informativity of context in recent speech input. When listeners experience context to be less informative in recent input, they adapt how they integrate subcategorical information with subsequent contextual cues that occur later in the speech signal. Under certain assumptions we discussed, this suggests that listeners can dynamically modulate whether they maintain subcategorical information during real-time processing, given the utility of such maintenance in recent input. Additionally, the present work provides further evidence that maintenance of subcategorical information beyond word boundaries is typical during spoken word recognition, rather than the exception: At least numerically, the majority of participants exhibited effects of both context *and* VOT (replicating Bicknell et al., 2025) and from the earliest moments of exposure (replicating

Bushong & Jaeger, 2019a). This stands in contrast to the view that listeners are unlikely to maintain any subcategorical information at distances significantly beyond the word boundary (Christiansen & Chater, 2016; Just & Carpenter, 1980).

Context of the Research

This research originated as part of the first author's PhD thesis on the maintenance of subcategorical information in spoken word recognition (Bushong, 2020). The overarching goal of this research is to understand the extent to which memory limitations shape spoken language understanding and whether any such limitations can be productively understood as a form of rational use of bounded cognitive resources.

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*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

Atienza, M., Cantero, J. L., & Gómez, C. M. (2000). Decay time of the auditory sensory memory trace during wakefulness and rem sleep. *Psychophysiology*, *37*(4), 485–493. https://doi.org/10.1017/s0048577200980697

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *LME4: Linear mixed-effects models using Eigen and S4* (R Package Version 1.1-4) [Computer software]. https://cran.r-project.org/web/packages/lme4/index.html

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLOS ONE*, *6*(5), Article e19812. https://doi.org/10.1371/journal.pone.0019812

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

Bicknell, K., Bushong, W., Tanenhaus, M. K., & Jaeger, T. F. (2025). Maintenance of subcategorical information during speech perception: Revisiting misunderstood limitations. *Journal of Memory and Language*, *140*, Article 104565. https://doi.org/10.1016/j.jml.2024.104565

Bottcher-Gandor, C., & Ullsperger, P. (1992). Mismatch negativity in event-related potentials to auditory stimuli as a function of varying interstimulus interval. *Psychophysiology*, *29*(5), 546–550. https://doi.org/10.1111/j.1469-8986.1992.tb02028.x

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, *32*(10), 1211–1228. https://doi.org/10.1080/23273798.2017.1325508

Burchill, Z. (2023). *The reliability of standard reading time analyses and understanding the nature of maintained information in speech processing* [Doctoral dissertation, University of Rochester].

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PLOS ONE*, *13*(8), Article e0199358. https://doi.org/10.1371/journal.pone.0199358

Bushong, W. (2020). *Maintenance of subcategorical information in spoken word recognition* [Doctoral dissertation, University of Rochester].

Bushong, W., & Jaeger, T. F. (2017). Maintenance of perceptual information in speech perception. *Proceedings of the thirty-ninth annual conference of the cognitive science society* (pp. 186–181).

Bushong, W., & Jaeger, T. F. (2019a). Dynamic re-weighting of acoustic and contextual cues in spoken word recognition. *The Journal of the Acoustical Society of America*, *146*(2), EL135–EL140. https://doi.org/10.1121/1.5119271

Bushong, W., & Jaeger, T. F. (2019b). Memory maintenance of gradient speech representations is mediated by their expected utility. *Proceedings of the forty-first annual conference of the cognitive science society* (pp. 1458–1463).

Bushong, W., & Jaeger, T. F. (2019c). Modeling long-distance cue integration in spoken word recognition. *Proceedings of the 2019 workshop on cognitive modeling and computational linguistics* (pp. 62–70).

Bushong, W., & Jaeger, T. F. (2024). *Changes in informativity of sentential context affects its integration with subcategorical information about preceding speech*. https://osf.io/cypg3

Caplan, S., Hafri, A., & Trueswell, J. (2021). Now you hear me, later you don't: The immediacy of linguistic computation and the representation of speech. *Psychological Science*, *32*(3), 410–423. https://doi.org/10.1177/0956797620968787

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

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, *30*(1), 234–250. https://doi.org/10.1016/0749-596X(91)90005-5

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin*, *96*(2), 341–370. https://doi.org/10.1037/0033-2909.96.2.341

Crowder, R. G. (1982). Decay of auditory memory in vowel discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(2), 153–162. https://doi.org/10.1037/0278-7393.8.2.153

Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science*, *19*(2), 121–126. https://doi.org/10.1177/0963721410364726

Falandays, J. B., Brown-Schmidt, S., & Toscano, J. C. (2020). Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language*, *112*, Article 104088. https://doi.org/10.1016/j.jml.2020.104088

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782. https://doi.org/10.1037/a0017196

Foulkes, P., & Hay, J. (2015) The emergence of sociophonetic structure. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 292–313). Blackwell.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325. https://doi.org/10.1016/0010-0277(78)90002-1

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873. https://doi.org/10.1002/sim.3107

Giovannone, N., & Theodore, R. M. (2021). Individual differences in lexical contributions to speech perception. *Journal of Speech, Language, and Hearing Research*, *64*(3), 707–724. https://doi.org/10.1044/2020_JSLHR-20-00283

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

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

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, *13*(1), Article 6606. https://doi.org/10.1038/s41467-022-34326-1

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

Hanulikova, A. (2022). The role of perceived ethnicity in speech processing: Insights from diverse populations and methods. *The Journal of the Acoustical Society of America*, *151*(4), Article A99. https://doi.org/10.1121/10.0010780

Hay, J., Walker, A., Sanchez, K., & Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PLOS ONE*, *14*(2), Article e0210793. https://doi.org/10.1371/journal.pone.0210793

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956. https://doi.org/10.1037/a0025641

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, *6*(8), 345–350. https://doi.org/10.1016/S1364-6613(02)01948-4

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Johnson, K. (1997). *The auditory/perceptual basis for speech segmentation* (Working Papers in Linguistics). The Ohio State University.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384. https://doi.org/10.1006/jpho.1999.0100

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, *34*(6), 769–786. https://doi.org/10.1080/23273798.2019.1582787

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221. https://doi.org/10.1121/1.380986

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311. https://doi.org/10.1111/tops.12086

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, Article e1. https://doi.org/10.1017/S0140525X1900061X

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70. https://doi.org/10.1016/j.cognition.2018.01.003

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, *44*(4), Article e12823. https://doi.org/10.1111/cogs.12823

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009).Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91. https://doi.org/10.1016/j.jml.2008.07.002

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62–85. https://doi.org/10.1177/0261927X99018001005

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–191. https://doi.org/10.1037/0033-295X.85.3.172

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Qian, T., & Jaeger, T. F. (2012). Cue effectiveness in communicatively efficient discourse production. *Cognitive Science*, *36*(7), 1312–1336. https://doi.org/10.1111/j.1551-6709.2012.01256.x

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(2), Article e1521. https://doi.org/10.1002/wcs.1521

Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Attention, Perception, & Psychophysics*, *75*(7), 1533–1546. https://doi.org/10.3758/s13414-013-0492-3

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x

Walker, A., & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, *2*(1), 219–237. https://doi.org/10.1515/labphon.2011.007

(*Appendix follows*)

Appendix

List of Stimuli

Tables A1 and A2 list the sentence items used in Experiments 1–3. Experiments 1 and 2 used only Items 1–12; Experiment 3 additionally utilized the neutral context Items 13–20 in order to decrease repetition of sentences in the low-informativity condition.

We note that even some of the “neutral” contexts are likely to nonnegligibly bias toward either *dent* or *tent*. To the extent that this is the case, this would bias *against* the hypothesis we tested—and found supported—in our experiments (since it reduces the difference between the informativity of low- and high-informativity exposure). There is a further difference between sentence types—in an effort to keep neutral sentences uninformative, they tended to be shorter on average than sentences with biasing context (by 3.2 words in items used for Experiments 1–2 and 3.25 words in Experiment 3; $p < .001$).

This meant that participants in the high-informativity group had to wait the equivalent of, on average, 3.2 words more before giving their response. As pointed out by a reviewer, it is possible that this helped participants become better at maintaining subcategorical information—not because of the higher informativity of the context, but rather because of the practice that these participants received in maintaining information. This struck us as an interesting possibility. It does, however, seem to be rather unlikely given the findings of additional generalized additive mixed-effects model analyses presented in the Supplemental Material (Section 2). Consider, for example, the left panels of Supplemental Figures S1, S3, and S5: Context effects in the high-informativity group always *decreased* during exposure. This is the opposite of what would be expected if the longer sentences during high-informativity allowed participants to practice the maintenance of subcategorical information. That said, future work should keep sentence length constant between exposure groups to more convincingly address this alternative interpretation of our findings.

Table A1
List of Stimulus Materials Used in Experiments 1–3

Item	Bias	Sentence	Alternate target
1a	*Tent*	Once the ?ent had been successfully set up, we made camp for the night.	Night/fight
1b	*Dent*	Once the ?ent had been repaired in the wall successfully, we were relieved.	Relieved/believed
1c	Neutral	Once the ?ent was made, we were done for the night.	Night/fight
2a	*Tent*	Since the ?ent the family got was quite sturdy, they ended up camping a lot.	Got/sought
2b	*Dent*	Since the ?ent the family got in the car was quite serious, it needed to be repaired.	Needed/heeded
2c	Neutral	Since the ?ent the family got was manageable, they were fine with it.	Got/fought
3a	*Tent*	Since the ?ent was hard to find in the trees, no one could see us.	Find/bind
3b	*Dent*	Since the ?ent was hard to find in the door, we never fixed it.	Find/mind
3c	Neutral	Since the ?ent was hard to find, nobody noticed it.	Find/mind
4a	*Tent*	After the ?ent Sue had found in the campgrounds collapsed, we went to a hotel.	Hotel/motel
4b	*Dent*	After the ?ent Sue had found in the teapot was noticed, we threw it away.	Threw/drew
4c	Neutral	After the ?ent was noticed, we continued on our way.	Way/day
5a	*Tent*	Since the ?ent was incredibly flimsy, it didn’t weather the storm.	Weather/tether
5b	*Dent*	Since the ?ent extended very deep, it was hard to fix.	Fix/mix
5c	Neutral	Since the ?ent had extensive damage, it was hard to fix.	Fix/mix
6a	*Tent*	Since the ?ent is so deep in the woods, it is difficult to find.	Deep/cheap
6b	*Dent*	Since the ?ent is so deep in the bike, the frame has rusted through.	Deep/cheap
6c	Neutral	Since the ?ent was so large, we made a note of it.	Note/boat
7a	*Tent*	Since the ?ent was removed from the camp, we were able to leave.	Leave/heave
7b	*Dent*	Since the ?ent was removed from the bucket, it has stopped leaking.	Leaking/reeking
7c	Neutral	Since the ?ent was removed, we were ready to go.	Go/tow
8a	*Tent*	When the ?ent was noticed in the forest, we stopped to rest.	Rest/nest
8b	*Dent*	When the ?ent was noticed in the fender, we sold the car.	Car/bar
8c	Neutral	When the ?ent was taken care of, we were ready to go.	Go/know
9a	*Tent*	Since the ?ent John had mistakenly bought was too small, the wedding guests had to stand in the rain.	Rain/lane
9b	*Dent*	Since the ?ent John had mistakenly caused was very large, he had to have it fixed.	Fixed/nixed
9c	Neutral	Since the ?ent looked fine, John left it alone.	Looked/cooked
10a	*Tent*	After seeing the ?ent yesterday on the hill, we chose to camp at the bottom.	Chose/rose
10b	*Dent*	After seeing the ?ent yesterday on the fender, we managed to fix it.	Fix/nix
10c	Neutral	After seeing the ?ent yesterday, we went home.	Went/sent
11a	*Tent*	After the ?ent was spotted on the peak by the mountaineers, they were relieved.	
11b	*Dent*	After the ?ent was spotted on the truck, the driver was worried.	Worried/hurried
11c	Neutral	After the ?ent was spotted by the woman, she went home.	She/he
12a	*Tent*	Because the ?ent that was found in John’s backpack had a few holes, he had to patch it.	Patch/match
12b	*Dent*	Because the ?ent that was found in John’s bumper was very small, he decided not to fix it.	Fix/mix
12c	Neutral	Because the ?ent wasn’t a big deal to John, he didn’t mind it.	Mind/find

(*Appendix continues*)

Table A2
List of Additional Neutral-Context Stimuli Used in Experiment 3

Item	Bias	Sentence	Alternate target
13d	Neutral	Although the ?ent was dealt with quickly, it was annoying.	Dealt/felt
14d	Neutral	Although the ?ent was easy to repair, we didn't have the right tools.	Tools/pools
15d	Neutral	Although the ?ent looked like it had been there for a while, it was new.	While/file
16d	Neutral	Although the ?ent seemed large, it really wasn't.	Seemed/deemed
17d	Neutral	Although the ?ent was ugly, it wasn't that bad.	Bad/rad
18d	Neutral	Because the ?ent was hard to see, we had to point it out.	Point/joint
19d	Neutral	Although the ?ent was well concealed, it was still noticed.	
20d	Neutral	When the ?ent was finally repaired, we were relieved.	Repaired/repelled

In Experiment 1, the target word that participants were asked to categorize was always *?ent* and the two response options shown on the screen were always *dent* and *tent*. In Experiments 2 and 3, participants were asked to categorize target words other than the *?ent* on half of all trials (regardless of exposure condition). These target words are listed in the "Alternate target" column of Tables A1 and A2. Not every sentence frame–condition combination used in the experiment had an alternate target word. We tried to keep the alternate words somewhat plausible within the sentence context, which was not possible with some of the sentence frames we used.

Received January 20, 2024
Revision received October 1, 2024
Accepted November 4, 2024 ■