Do they know it's Christmash? Lexical knowledge directly impacts speech perception

Sahil Luthra[1], Anne Marie Crinnion[2], David Saltzman[2] & James S. Magnuson[2,3,4]

[1]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, United States
[2]Department of Psychological Sciences, University of Connecticut, Storrs, CT, United States
[3]BCBL: Basque Center on Cognition, Brain and Language, Donostia-San Sebastián, Spain
[4]Ikerbasque: Basque Foundation for Science, Bilbao, Spain

Corresponding author:
Sahil Luthra
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
sahilluthra@cmu.edu

Author Note

All data, analysis scripts and figures are available at https://osf.io/mdn8w/. This research was supported in part by U.S. National Science Foundation grants BCS-PAC 1754284, BCS-PAC 2043903, and NRT 1747486 (PI: JSM). This research was also supported in part by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and through project PID2020-119131GB-I00 (BLIS). SL was supported by NIH NRSA F32DC020625, AMC by NIH NRSA F31DC021372 and DS by NIH NRSA F31DC019873. AMC was also supported by NIH T32 DC017703 (E. Myers and I-M. Eigsti, PIs).

Abstract

We recently reported strong, replicable (i.e., replicated) evidence for *lexically-mediated compensation for coarticulation* (LCfC; Luthra et al., 2021), whereby lexical knowledge influences a pre-lexical process. Critically, evidence for LCfC provides robust support for *interactive* models of cognition that include *top-down feedback* and is inconsistent with *autonomous* models that allow only feedforward processing. McQueen, Jesse and Mitterer (2023) offer five counter-arguments against our interpretation; we respond to each of those arguments here and conclude that top-down feedback provides the most parsimonious explanation of extant data.

Introduction

A core debate in cognitive science centers on whether high-level knowledge directly shapes perception or merely influences post-perceptual interpretations (e.g., Firestone & Scholl, 2016; Lupyan, 2015; Magnuson et al., 2018; Norris, McQueen & Cutler, 2016). Norris, McQueen and Cutler (2000) argued that apparent top-down effects in spoken word recognition reflect post-perceptual integration, consistent with *autonomous* theoretical accounts that preclude feedback.

Elman and McClelland (1988) devised a critical test case that cannot be explained as post-perceptual integration. Their paradigm combines phoneme restoration (e.g., listeners identify an ambiguous fricative as /s/ given the frame "Christma_" but as /ʃ/ ["sh"] given "fooli_"; Ganong, 1980) and a prelexical process known as compensation for coarticulation (CfC). In CfC, the perception of a place ambiguity is influenced by the place of articulation of a preceding segment (e.g., an ambiguous step from a "tapes"-"capes" continuum is more likely to be heard as /t/ after /ʃ/ and as /k/ following /s/; Mann & Repp, 1981). Elman and McClelland found that a fricative restored as /s/ or /ʃ/ (front place of articulation vs. back) by lexical context ("Christma_" vs. "fooli_") could drive CfC on a subsequent /t/-/k/ ambiguity (indicating that the final phoneme was genuinely restored, since it was able to influence the *perception* of the following /t/-/k/ ambiguity). This *lexically-mediated compensation for coarticulation* (LCfC) would constitute strong support for interactive models and could not be explained by autonomous accounts. However, LCfC has not been consistently observed, with some positive reports (Elman & McClelland, 1988; Magnuson et al., 2003a; Samuel & Pitt, 2003) and some negative reports (Pitt & McQueen, 1998; McQueen, Jesse & Norris, 2009).

Luthra et al. (2021; cf. Samuel & Pitt, 2003) noted that few LCfC studies pretested materials to confirm that they could independently drive lexical phoneme restoration (Ganong) and CfC effects. Materials that cannot drive component effects separately will not be capable of driving LCfC. With rigorously pretested materials, we observed robust LCfC in a well-powered, pre-registered study and an independent replication sample. McQueen, Jesse and Mitterer (2023) argue that our findings should not be taken as evidence for top-down processing; here, we consider their five arguments.

1. Accounting for null results

McQueen et al. (2023) write that "Before Luthra et al. (2021) can conclude in favor of top-down processing based on their data from the LCfC paradigm… they need to offer a convincing explanation for other data from the paradigm that contradict their account" (p. 3). They focus particularly on null results from McQueen et al. (2009) and argue that Bayes Factor analyses favor the null hypothesis. We disagree with three aspects of their analyses.

First, their Bayes Factor analyses test for LCfC under the assumption that the lexical restoration effect size for context items should predict the LCfC effect size. While interactive theories do

assume that these effects originate from the same source (lexical knowledge), there are many reasons why these effects might not correlate. For instance, Ganong and CfC effects would not be expected to correlate if there is a ceiling on the size of either effect or if within-subject reliability in the measurement of either effect is poor. Furthermore, there is no guarantee that any given experimental paradigm will be sensitive enough to detect subtle gradations of an effect, especially if we are near the limit of what the paradigm can detect. Thus, *the hypothesis being evaluated by the Bayes Factor analyses is not necessarily reflective of the interactive view and therefore does not constitute a valid test of feedback*.

Second, McQueen et al. (2023) applied their Bayes Factor analyses to three ambiguous steps from a phonetic continuum. As they acknowledge, individual Bayes Factor analyses are inconclusive for two of three continuum steps and provide only moderate support for the null hypothesis in the third case. (By convention, Bayes Factors between ⅓ and 3 are considered to be inconclusive, whereas those that are less than or greater than those bounds are considered to be evidence in favor of the null or in favor of the alternative, respectively; see Dienes, 2014.) The authors also conducted a combined Bayes Factor analysis by multiplying individual Bayes Factors together, providing a rationale for this approach in their Supplementary Materials. However, multiplying Bayes Factors is problematic when data are drawn from the same distribution (Rouder & Morey, 2011); critically, the data from the three continuum steps are drawn from the same distribution, since each participant heard all three (i.e., Continuum Step was a within-subjects factor). Strikingly, if one appropriately pools the ambiguous steps from the McQueen et al. data (rather than treating each as separate, independent samples), the result is inconclusive (BF = 0.45; see analysis scripts at https://osf.io/mdn8w/).

Third, we are unsure what McQueen et al. infer is proven by their analyses. To take their BF analyses of a single experiment as evidence that the null hypothesis is more likely than LCfC requires ignoring all positive results.

Furthermore, insisting we explain all null effects is not an appropriate burden of proof; there are myriad reasons why a study could fail to see a significant effect. For example, consider the data from McQueen et al. (2009), who observed Ganong and CfC effects in one set of trials but failed to observe LCfC on a separate set of trials. Critically, this design required the use of a 4-alternative forced choice (4AFC) task, wherein listeners were required to categorize both the final segment of the context item and the first segment of the target item. It is possible that this cognitively demanding 4AFC task obscured potential LCfC effects, especially since previous work suggests that some perceptual effects in speech processing may be attenuated when task demands are heightened; for example, increasing cognitive load can attenuate cross-modal phonetic recalibration effects driven by visual (lipreading) information (Jesse & Kaplan, 2019), and shifting attentional resources away from the speech signal can extinguish the influence of lexical knowledge on phonetic retuning (Samuel, 2016). In addition to simple

complexity/demand issues, note that McQueen and colleagues in various papers have argued that the Ganong effect results from post-perceptual bias, while accepting that CfC is a perceptual-level effect. On that logic (with which we do not agree, as the interactive account describes Ganong effects as perceptual results of top-down feedback), the 4AFC forces participants to integrate a post-perceptual decision with what is normally a perceptual decision, contaminating the putatively online decision component with the putatively post-perceptual decision component.

For these reasons, we argue that it is preferable to establish Ganong and CfC effects via isolated pretesting (Luthra et al., 2021), allowing for the use of a simpler 2AFC task that minimizes task demands. Furthermore, a key implication of our recent paper (Luthra et al., 2021) is that poor stimulus construction may be to blame when component Ganong and CfC effects are not first established with pretesting. We argue that the more appropriate challenge is for proponents of the autonomous perspective to explain positive LCfC results – especially from studies that test for robust, independent Ganong and CfC effects prior to testing for LCfC.

2. Transitional probabilities could somehow explain results

McQueen et al. (2023) argue that phoneme-to-phoneme transitional probabilities (TPs) might explain (at least some) LCfC effects, citing evidence that TPs can influence CfC (Pitt & McQueen, 1998). However, evidence that TPs can influence CfC is not evidence against lexical influences on CfC; it simply demonstrates that TP-mediated CfC is also possible.

In the past (McQueen et al., 2009; Pitt & McQueen, 1998; and in reviews of this response), the authors have also made this argument based on computational simulations by Norris (1993), who showed that a Simple Recurrent Network (SRN; Elman, 1990, 1991) can simulate LCfC. Norris trained an SRN to map 11-feature phonetic inputs to separate outputs representing the current word and phoneme. The critical items were 12 CVC words. Phonetic inputs were adjusted to reflect CfC (shifting the features for adjacent phonemes with different places of articulation toward each other). The critical items built in a single TP contingency: the final C was predicted by the initial CV. After training, the SRN demonstrated LCfC on the critical items.

However, for this to support autonomous architectures, it is necessary that SRNs are purely feedforward models. This is not the case. Norris asserts that there is no form of feedback in an SRN because the recurrent connections are from hidden nodes to hidden nodes with a time delay of 1 step, and therefore constitute lateral connections, rather than feedback. But feedforward networks and recurrent networks are fundamentally different. The input at time *t* to the hidden layer is both the current input pattern but also the states of the hidden nodes at the previous time step (typically copied to a context layer); those states are themselves a combination of the previous input the hidden states two time steps back with the input from one time step back, and so on. This means that the input to the hidden layer (the point where inputs first impinge on the

SRN) *includes information that results from a transformation carried out internally within the model (context x hidden connections*, which again are products of mixtures of inputs times input-to-hidden weights combined with context states [hidden at previous time step] multiplied by context-to-hidden [or hidden-to-hidden] weights); this is top-down interaction. The hidden nodes do not "know" which aspects of their input are external (from the actual input nodes) and which are internal (from hidden nodes); the external inputs are mixed with internal information, precluding veridical input encoding. More formally, recurrent networks are cyclic graphs (i.e., they contain loops, in contrast to feedforward networks, which are a acyclic because they have no loops; Prince, 2024), and computer science treatments of this distinction routinely describe recurrent networks as dependent on feedback of their own prior, transformed states (Jurafsky & Martin, forthcoming; Prince, 2024). We make this case in more detail in Magnuson and Luthra (submitted).

Strikingly, consider how Norris et al. (2016) characterize what they see as flaws entailed by adding feedback: "the problem here is that the activation generated by the input is being reused multiple times, and amplified each time" (p. 5), potentially leading to hallucinations. Thus, to them, the critical problem is feedback mixes bottom-up inputs with top-down information, precluding veridical perception. The same is true of SRNs, since inputs are inextricably mixed with model states from the previous timestep, which includes combined transformations from the previous step's input and context states (recursively over preceding time steps). Redescribing the SRN as having time-delayed connections between hidden units (as Norris [1993] does) does not resolve this; the math remains the same, and it is a fact that inputs are immediately mixed with model internal transformations of the previous time step(s), making SRNs a form of interactive model.

McQueen et al. (2023) have not offered a computationally specific proposal for how TPs could explain apparent lexical effects; critically, a testable hypothesis would require identifying the order of transitional probabilities (e.g., diphones, triphones, or some specific combination of multiple orders) that would underlie this effect. Without a concrete proposal based on TPs, it is impossible to test whether potential LCfC effects truly reflect a lexical influence or simply TPs; any putatively lexical effect could simply reflect an *n*-phone influence, where *n* is the TP order that is consistent with the specific item under examination (sometimes diphone, sometimes triphone, …). Without a computationally specific explanation of how TPs might explain lexical influences, there simply is no empirical way to engage with the critique, which appeals to diphone probabilities when those are consistent with the explanation (Pitt & McQueen, 1998) but to triphone probabilities when those turn out to be consistent with the explanation (e.g., McQueen et al., 2023 appeal to triphone TPs as well as "higher-order TP biases", p. 4). Nevertheless, we have attempted to test this proposal as best we can, by assessing whether any of multiple possible orders of TP could coherently account for positive LCfC results. We have shown in multiple analyses that there is no identifiable order of *n*-phone or set of *n*-phones that

provide a comprehensive, item-specific explanation for positive results in LCfC (Luthra et al., 2021; Magnuson et al., 2003b). Instead, the predictive context most often resolves to word length minus one — i.e., lexical context.

Additionally, it is worth emphasizing that TPs do not necessarily have a bottom-up basis. In the TRACE model (McClelland & Elman, 1986), for example, TP effects emerge via combined impacts of top-down lexical feedback and lateral inhibition (another emergent property of interactive activation that does not require a separate mechanism to be invoked). Furthermore, as justified above, simulations of LCfC with SRNs also depend on mixing current bottom-up inputs and (top-down) previous network transformations.

3. Learning over the course of the experiment

McQueen et al. (2023) argue that listeners could learn over the course of the experiment that the ambiguous ending of the context item is always the lexically consistent one. Here, they appeal to McQueen et al. (2009), who showed that the nature of practice trials can influence the emergence of putative LCfC effects; for this reason, we did not include practice trials in Luthra et al. (2021). Critically, if it were the case that the results observed by Luthra et al. were due to learning over the course of the experiment, we would expect the putatively lexical effect to only emerge after a countable number of initial trials. McQueen et al. conducted an analysis with trial as a factor but found no evidence that LCfC was influenced by trial. They then argue that the learning may have occurred so rapidly as to be undetectable over trials. If learning is proposed to be so fast that it is undetectable, it seems there is no scientific avenue to measure (let alone test) their learning hypothesis.

In motivating their argument for experiment-induced learning, McQueen et al. (2023) also appeal to studies of perceptual learning. For example, it has been documented that lexically guided perceptual learning can be obtained with as few as 10 exposures to critical ambiguous stimuli (Kraljic & Samuel, 2007) and that such learning scales with dosage of exposure to the critical stimuli (Cummings & Theodore, 2023); in these studies, the (sometimes small number of) critical stimuli are presented during an initial exposure phase, and learning is assayed during a separate test phase with a relatively large number of trials. Note that it would be exceedingly difficult (and perhaps impossible) to conduct a study of dosage with the LCfC paradigm, since manipulating dosage (i.e., the number of exposures to the ambiguous context items) necessarily also changes the number of trials used to estimate the size of the LCfC effect. Because LCfC is measured via a categorization function (see Figure 1), an experimenter must obtain several measurements at each continuum step for each context. If experiment-induced learning could in principle occur with as few as one exposure to an ambiguous context stimulus, it is therefore not clear how one could assay the impact of exposure dosage in the LCfC paradigm, short of conducting a study with only one observation per participant (introducing severe concerns about statistical power, or requiring massive numbers of participants).

The authors also write that "What is required to prevent this kind of learning is exposure to unambiguous tokens of both interpretations of the ambiguous sound (e.g., Christma[s/ʃ] with Christmas and Christmash, in equal proportions; fooli[s/ʃ] with foolis and foolish)" (p. 5). In our study, participants receive equal proportions of both interpretations in that they receive zero tokens of each. Additionally, it is incredibly unlikely that a listener would hear "Christmash" in equal proportion to "Christmas" in real-world listening conditions, making the proposed experiment problematic in terms of ecological validity, as non-ecological distributions can influence effect sizes in phonetic categorization studies (Bushong & Jaeger, 2019). Specifically, Bushong and Jaeger argue that unecological distributions (in particular, including clear nonword tokens) distort the normal influence of top-down knowledge.

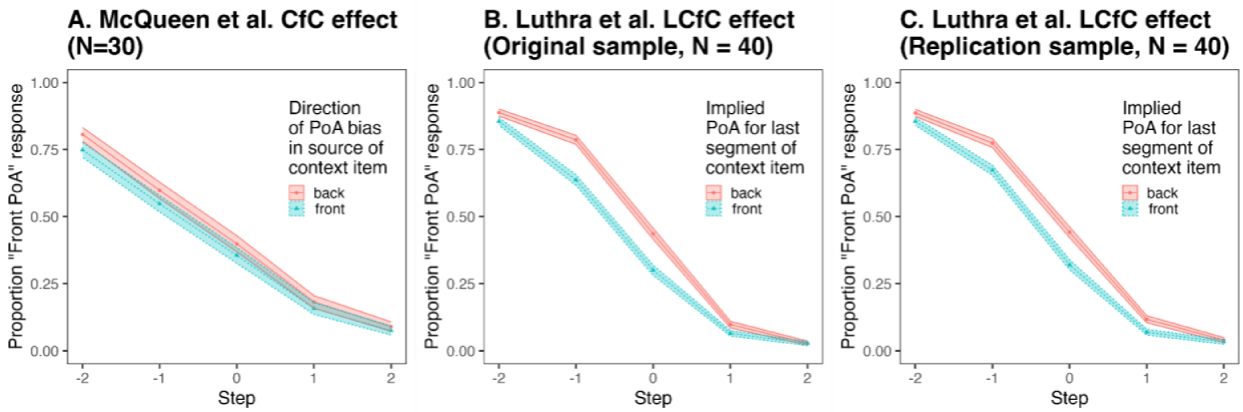

Figure 1. McQueen et al. (2023) argue that there was a confound in some context stimuli used by Luthra et al. (2021), in that the maximally ambiguous token from the *isolate*/**isolake* continuum was slightly biased toward a front place of articulation and the maximally ambiguous token from the *maniac*/**maniat* continuum was slightly biased toward a back place of articulation. They conducted a cross-splicing experiment and found that in the absence of lexical information, these ambiguous blends could drive a small but reliable CfC effect in the same direction as the observed LCfC effect. However, the size of their CfC effect (panel A) is substantially smaller than the LCfC effect observed by Luthra et al. (panel B: original sample; panel C: replication experiment), which we take as evidence that the lexical influence substantially drove our observed and replicated results. Note also that the other two contexts (*pocketful*/**pocketfur* and *questionnaire*/**questionnaile)* were biased *against* the lexically consistent ending. Ribbons indicate within-subject 95% confidence intervals, estimated using the *summarySEwithin* function in the "Rmisc" package (Hope, 2022) in R (R Core Team, 2022).

4. Acoustic differences between stimuli

McQueen et al. (2023) argue that it is necessary to have identical acoustics for context-final segments (e.g., identical ending segments for *mania[t/k]* and *isola[t/k]*). While there are certainly advantages for comparing acoustically identical ambiguous stimuli, the advantage of our

approach is that each context's ambiguous stimulus is the maximally ambiguous morph between naturally produced tokens. McQueen et al. note our t/k context stimuli have a slight acoustic bias toward the lexical endpoint. They provide some empirical evidence that even when lexical information is removed, these context-final ambiguous stimuli can drive small but significant CfC effects in the direction of our LCfC effects. However, their CfC effect (a 3.8% difference between conditions, averaging across continuum steps) is substantially smaller than our LCfC effects (a 7.1% difference between conditions in our original sample and a 6.2% difference in our replication sample, though this difference is especially large at intermediate continuum steps; see Figure 1). We therefore argue that the majority of the effect observed by Luthra et al. (2021) is attributable to top-down lexical feedback.

McQueen et al. (2023) also describe an acoustic difference in the l/r context stimuli used by Luthra et al. (2021) — specifically, that there is devoicing in one context (*pocketful*-**pocketfur*) but not the other (*questionnaire*-**questionnaile*) — and write that they are unable to conduct a cross-splicing experiment because of the strong coarticulation from the vowel to the word-final segment. They suggest that this acoustic difference could contribute to a spurious basis for our effects. However, one can characterize the acoustic bias in the context stimuli by trimming them to form contexts that are ambiguous between two nonwords (e.g., trimming the ambiguous *pocketful*/**pocketfur* stimulus to be **ul*/**ur*) and examining which endpoints they are biased toward. In this way, Luthra et al. (2021) determined that the l/r contexts they used actually have slight acoustic biases *away* from the lexical endpoints (see Supplementary Materials). The tested bias (Figure 1) and the untested bias are in opposite directions, and thus may cancel each other out. We take this as additional evidence that acoustic differences between the context stimuli cannot account for the effects observed by Luthra et al.

5. Unspecified interactions

McQueen et al. (2023) argue that our effects could be "the result of acoustic effects, TP or experiment-induced biases, or their combination, and, when combined, those effects could have amplified each other" (p. 10). However, without a fully specified (e.g., computationally implemented) mechanistic explanation of how transitional probabilities and experiment-induced biases could explain the current data, and evidence or explanation for how they might interact, it is not possible to quantify (and thus test or falsify) the authors' suggestion that these factors could conspire, in some unspecified way, to drive our robust and replicated results.

Concluding remarks

The question of whether there are top-down effects in speech perception has substantial implications for the cognitive and neural sciences; it is no surprise that this question has inspired fervent debate. McQueen et al. (2023) raise interesting questions and identify (minor) flaws in a subset of our materials but fail to provide a comprehensive refutation of the growing body of positive LCfC results. Their Bayes Factor analyses only favor the null hypothesis in a minority

of published results (one experiment), and even then depend on parameter choices that favor that conclusion. After 20 years, they have yet to provide an account of how transitional probabilities could explain apparent lexical effects in the specific items that have yielded positive LCfC results. We would argue that their appeal to within-experiment learning is unfalsifiable, specifically their claim that learning might take place so early in the experiment as to be undetectable with trial-level analyses (McQueen et al., 2023, p. 5). Among the acoustic differences they identify in our stimuli, they only tested one that affects a subset of context items, but their experimental test of whether the slight bias in those items could drive our LCfC effect yields an effect considerably smaller than our replicated LCfC result. The kitchen-sink appeal to these four issues conspiring in an unknown way to produce systematically positive LCfC results is unconvincing. Interaction, on the other hand, provides a coherent and parsimonious account.

Although we disagree with the conclusions of McQueen et al., we are grateful for their careful scrutiny of our work and applaud them for making their analysis materials openly available; these open science practices will be key for resolving the debate over how listeners integrate high-level knowledge with sensory input.

References

Bushong, W., & Jaeger, T. F. (2019). Dynamic re-weighting of acoustic and contextual cues in spoken word recognition. *The Journal of the Acoustical Society of America*, *146*(2), EL135-EL140.

Cummings, S. N., & Theodore, R. M. (2023). Hearing is believing: Lexically guided perceptual learning is graded to reflect the quantity of evidence in speech input. *Cognition*, *235*, 105404.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179-211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195-225.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27(2), 143–165.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, 39, 1–77.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. Journal ofExperimental Psychology: Human Perception and Performance, 6(1), 110–125.

Hope RM (2022). Rmisc: Ryan Miscellaneous. R package version 1.5.1, <https://CRAN.R-project.org/package=Rmisc>.

Jesse, A., & Kaplan, E. (2019). Attentional resources contribute to the perceptual learning of talker idiosyncrasies in audiovisual speech. *Attention, Perception, & Psychophysics, 81*(4), 1006-1019.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1-15.

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*, 6(4), 547–569.

Luthra, S., Peraza-Santiago, G., Beeson, K. N., Saltzman, D., Crinnion, A. M., & Magnuson, J. S. (2021). Robust lexically mediated compensation for coarticulation: Christmash time is here again. *Cognitive Science*, *45*(4), e12962.

Magnuson, J. S. & Luthra, S. Simple Recurrent Networks are interactive. Submitted.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003a). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science*, 27(2), 285–298.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003b). Lexical effects on compensation for coarticulation: A tale of two systems? *Cognitive Science*, 27(5), 801–805.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. D. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, 9, 1–18.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69*(2), 548–558.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1-86.

McQueen, J. M., Jesse, A., & Mitterer, H. (2023). Lexically-mediated compensation for coarticulation still as elusive as a white Christmash. *Cognitive Science, 47,* e13342.

McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical–prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes?. *Journal of Memory and Language*, *61*(1), 1-18.

Norris, D. (1993). Must connectionist models be interactive? In *Cognitive Models of Speech Processing: The Second Sperlonga Meeting* (p. 211-234). Psychology Press.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(3), 299-325.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204-238.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, 31(1), 4–18.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon?. *Journal of Memory and Language*, *39*(3), 347-370.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682-689.

Samuel, A. G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive Psychology, 88*, 88-114.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language*, *48*(2), 416-434.

Supplementary Materials

McQueen, Jesse and Mitterer (2023) assert that Luthra et al. (2021) used problematic stimuli in testing for lexically mediated compensation for coarticulation (LCfC). They highlight the acoustic differences in our context stimuli, including a difference in devoicing between the *pocketfu?* and *questionnai?* contexts. Notably, data from Luthra et al. provide evidence that this acoustic difference leads to biases *against* lexical context, not toward.

In one of their pilot experiments, Luthra et al. (2021) presented one group of listeners with word-nonword continua (e.g., *pocketful*-**pocketfur*) and a separate group of listeners with nonword-nonword continua generated by trimming the word-nonword continua (e.g., **ul*-**ur*). For a context item to be included in subsequent experiments, there had to exist at least one step where participants who heard the word-nonword continuum made more lexically consistent responses compared to participants who heard the associated nonword-nonword continuum; that is, there had to be one step where a lexical (Ganong, 1980) effect was observed. This step was identified as the maximally ambiguous step, and that continuum step was used for the main LCfC experiments.

Critically, we can also assess the bias of the context stimuli by looking at responses to the nonword-nonword continua at the most ambiguous step (Figure S1, circled points). In this way, we can see that the *pocketful*-**pocketfur* continuum is biased toward the lexically inconsistent /r/ endpoint (as participants only labeled this token as /l/ 38% of the time) and the *questionnaire*-**questionnaile* continuum is biased toward the lexically inconsistent /l/ endpoint (as participants labeled this token as /l/ 66% of the time). While this pilot experiment did not assess the effect of these context stimuli on labeling of the the subsequent target continua, it does provide evidence that these context stimuli had acoustic biases that were inconsistent with lexical knowledge, making it unlikely that the LCfC effect observed in the main experiments was driven by the acoustic differences between these stimuli.

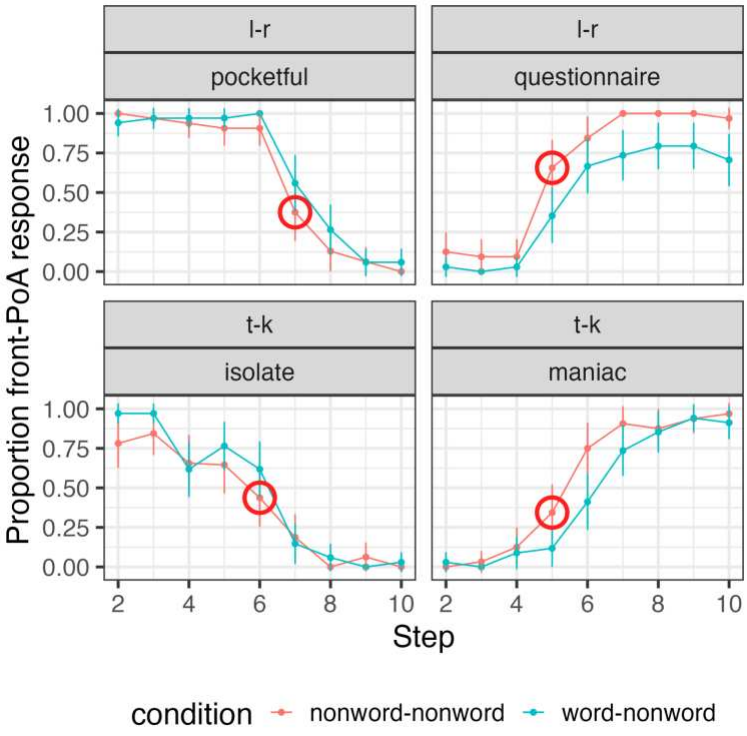

Figure S1. Results from the context item pilot experiment from Luthra et al. (2021). Circles indicate the front-rate for a nonword-nonword context at the "maximally ambiguous" step of each continuum. *Pocketful* is r-biased and *questionnaire* is l-biased (so the steps selected build in a bias **against** lexical context). *Isolate* and *maniac* are both k-biased, though *isolate* is less k-biased than *maniac*, making the baseline bias of these items consistent with lexical context.