Lexical Knowledge Enhances Consistency in Speech Categorization

Sita Carraturo [a*], Hyoju Kim [b], Ethan Kutlu [b, d], Bob McMurray [b, c, d]

[a] Department of Otolaryngology, University of Iowa, Iowa City, IA 52242, USA
[b] Department of Psychological & Brain Sciences, University of Iowa, Iowa City, IA 52242, USA
[c] Department of Linguistics, University of Iowa, Iowa City, IA 52242, USA
[d] Department of Communication Sciences & Disorders, University of Iowa, Iowa City, IA 52242, USA

* Corresponding author: Sita Carraturo, Department of Psychological & Brain Sciences, University of Iowa, PBSB 255E, Iowa City, IA, 52242, United States of America.

E-mail addresses: sita-carraturo@uiowa.edu (S. Carraturo), hyoju-kim@uiowa.edu (H. Kim), ethan-kutlu@uiowa.edu (E. Kutlu), bob-mcmurray@uiowa.edu (B. McMurray).

Acknowledgments

Part of the data presented here was presented at the *188th Meeting of the Acoustical Society of America* in New Orleans, Louisiana. This work was supported by the National Science Foundation under grant number BCS–2444664 to BM.

Declaration of Interest

The authors declare that they have no competing interests.

Data Availability

The registration document, data, and code for the present study are available at an Open Science Framework site for this project (https://osf.io/5b26j/).

CRediT author statement

Sita Carraturo: Methodology, Formal analysis, Writing – Original Draft, Visualization. **Hyoju Kim**: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft, Visualization. **Ethan Kutlu**: Conceptualization, Writing – Review & Editing. **Bob McMurray**: Conceptualization, Methodology, Sofware, Resources, Writing – Original Draft, Supervision, Project administration, Funding acquisition.

Word count: 4237

Abstract

Speech categorization is a gateway for downstream language processes. Recent work using the Visual Analog Scaling (VAS) task underscores the critical role of categorization consistency (trial-by-trial response variability around the mean response function) as a critical predictor of real-world outcomes such as language and reading abilities. Yet, the mechanisms that contribute to categorization consistency remain unknown. One hypothesis is that higher-level linguistic factors, such as lexical knowledge, may stabilize the percept by cleaning up lower-level perceptual noise. The first aim of this study was to test this hypothesis by examining whether categorization consistency is modulated by lexicality (word vs. nonword). Forty-eight adult American English listeners completed a VAS task involving both word (e.g., *batch-patch*) and matched nonword (e.g., *bazg-pazg*) continua. Listeners' categorization consistency for the word continua was significantly higher than for the nonword continua. This suggests that categorization consistency is, indeed, affected by higher-level linguistic factors. The second aim was to investigate whether individuals' broader language abilities influence their reliance on lexical information during speech categorization. Although individuals with greater language ability showed more consistent categorization, language scores did not modulate the lexical effect on categorization consistency. Together, these findings demonstrate the roles of top-down knowledge and language knowledge in stabilizing speech categorization.

Keywords: Speech categorization, Visual analog scaling, Lexical feedback, Categorization consistency, Individual differences

1. Introduction

Understanding spoken language is contingent on speech perception, the process of mapping variable speech signals to stored phonological or lexical representations. Decades of psycholinguistic research has sought to uncover the nature of the categories that people use to represent speech (such as /b/ and /p/), and how listeners identify these categories despite the inherent acoustic variability in speech (Clayards et al., 2008; Goldinger, 1998; Holt & Lotto, 2010; Liberman et al., 1967; McMurray, 2022; McMurray & Jongman, 2011; Stilp, 2020).

Historically, such work emphasized broad characterizations that apply across listeners, but more recently, research has sought to characterize individual variation in speech categorization (Fuhrmeister & Myers, 2021; Kapnoula et al., 2017; Kim, Klein-Packard, et al., 2025; Kim et al., 2024; Kutlu et al., 2024; Ou et al., 2021). This new focus has led to new questions about the underlying mechanisms that lead to those differences. Early individual differences work asked classic questions about categorical perception, category learning, and cue integration, but from an individual-differences perspective (e.g., Clayards, 2018; Kapnoula et al., 2017; Kong & Edwards, 2016; McHaney et al., 2021), essentially extending classic theories and debates to individual differences. A critical recent finding, however, is that one of the most important indices that differs across people is something entirely new: how consistently a listener arrives at the same percept across repeated trials—their *categorization consistency* (Honda et al., 2024; Kim, Klein-Packard, et al., 2025; Kim, McMurray, et al., 2025; Myers et al., 2024).

The importance of categorization consistency is demonstrated by its relationship to real-world outcomes. Among children, higher categorization consistency is strongly linked to better oral language and word reading (Kim, Klein-Packard, et al., 2025). In adults, it predicts about 30% of the variance in language ability, even after controlling for consistency in a non-speech visual categorization task (Kim et al., 2024). Greater consistency also predicts better performance on speech-in-noise tasks (Myers et al., 2024) and more accurate discrimination of speech sounds in second/foreign languages (Honda et al., 2024).

These findings raise a key question: where does categorization consistency (and individual variation therein) come from? Both bottom-up and top-down mechanisms are likely to contribute. Regarding bottom-up mechanisms, variability in low-level auditory encoding may be one source. For instance, children with more stable auditory brainstem responses to speech stimuli exhibit better reading ability (Hornickel & Kraus, 2013; Neef et al., 2017). This common link—between consistency of a perceptual response and reading—raises the possibility of similar mechanisms at play and supports the idea that noise at the earliest levels of auditory processing can propagate upward, influencing how consistently listeners categorize ambiguous speech sounds.

At the same time, top-down mechanisms could also contribute. Higher-level factors like lexical knowledge (Luthra et al., 2021), talker familiarity (McMurray & Jongman, 2016), and sublexical regularities (Newman et al., 1997; Pitt & McQueen, 1998) could exert feedback influences on perceptual processing and could help stabilize noisy or ambiguous input (Luthra et al., 2024; Magnuson et al., 2024). This connects to a long-running debate over how higher-level information interacts with low-level input (Magnuson, 2025; Magnuson & Luthra, 2024; Norris et al., 2000). We argue, however, that the feedback mechanisms usually proposed in this literature are distinct from what is needed to achieve greater categorization consistency.

Traditionally, feedback from higher-level representations to lower-level processing has been thought to serve one of two roles: (1) to bias the percept toward expectations (i.e., the

Ganong Effect; Ganong, 1980), or (2) to generate a difference between percept and expectations (Blank & Davis, 2016; McMurray & Jongman, 2011; and see Lupyan, 2017 for an example in vision). Empirical evidence for such feedback has focused on detecting bias in phoneme categorization (e.g., a shift in the boundary; Elman & McClelland, 1988; Norris et al., 2000) or accuracy differences as a function of context (e.g., McMurray & Jongman, 2016). Under these accounts, feedback *alters* existing representations to better account for broader context. This is largely studied through its effects on the average response across trials (as these effects are presumed to be equally operative over time).

In contrast to these traditional views, we propose an alternative function for top-down feedback—what we refer to as Auditory/Phonological Clean-Up (APCU). According to this hypothesis, feedback from higher-level representations (e.g., lexicality) does not merely bias perceptual outcomes, but actively contributes to stabilizing inherent noise in the perceptual system such that the percept is more consistently aligned with the signal. Lexical knowledge may act as a scaffold that reduces variability in the perceptual interpretation of speech sounds, leading to more consistent categorization across repeated presentations of the same input. This proposal aligns with recent computational modeling work using the TRACE model (Magnuson et al., 2024), which demonstrates that lexical feedback can "sharpen" noisy input representations.

Robust evidence for the APCU hypothesis would require clear evidence of feedback as the locus of the clean-up. As the long-running debate over lexical feedback attests, direct (or indirect) evidence of alterations to sublexical representations is an inherently difficult enterprise (Firestone & Scholl, 2016; Norris et al., 2000), requiring clever chained effects (Elman & McClelland, 1988; Samuel, 2001), or more direct access to sublexical representations via neuroscience (e.g., Getz & Toscano, 2019; Noe & Fischer-Baum, 2020; Sarrett et al., 2020). However, before attempting a more complex paradigm, this study seeks to first establish the viability of the APCU hypothesis by asking whether lexicality contributes to increased categorization consistency at all. By itself, such a finding would be consistent with both a feedback-induced locus and with other post-perceptual mechanisms. However, an absence of an effect would rule out the APCU hypothesis entirely.

Thus, as our first aim, we compared categorization consistency using speech continua that spanned two words (e.g., *bond/pond*) or two closely matched nonwords (*bonf/ponf*). Critically, the manipulated acoustic cue was identical in both conditions, making the broader lexical context the only relevant factor. Categorization consistency was measured using the Visual Analog Scaling (VAS) task (Apfelbaum et al., 2022; Massaro & Cohen, 1983; Munson et al., 2017), in which listeners heard tokens from a continuum (e.g., *beach-peach*) and rated each token along a continuous analog scale. This provides a fine-grained index of perceptual gradiency (slope of the categorization function), which can speak to any effects that operate across trials on the average representation. More importantly, once the slope is estimated, we can precisely assess trial-by-trial variability around the mean function—an index of categorization consistency.

Our second (exploratory) aim was to clarify the relationship between categorization consistency and broader language abilities. We build on recent work showing that categorization consistency predicts both adults' and children's language ability (Kim, Klein-Packard, et al., 2025; Kim et al., 2024). However, these studies only used word-word continua, leaving it unclear whether variation in cleanup or variation in the inherent noise is more important. Thus, we included the same language measures employed by Kim et al. (2024) and predicted an interaction with a stronger effect of language for words than nonwords. However, the exact nature of this interaction is unclear. One possibility is that individuals who show a

greater effect of word-nonword continua may demonstrate better language abilities, as lexical feedback is known to enhance word recognition speed and accuracy (Magnuson et al., 2024; Magnuson et al., 2018). On the other hand, greater reliance on higher-level information may be a compensatory mechanism and, therefore, linked to poorer language abilities.

2. Methods

2.1. Open Science Practices

This study was pre-registered. Stimuli, registration, data, and analysis code are available at https://osf.io/5b26j/?view_only=732dfad4b90f417e9b9e387db11d62ad.

2.2. Participants

Participants were recruited via Prolific (www.prolific.com). All subjects provided informed consent in accordance with the University of [redacted for review] Institutional Review Board and received compensation ($12/hour). Fifty-eight participants initially completed the tasks. After exclusions (detailed below), the final sample included 48 native speakers of American English (22 female, M_{age} = 30.5 years, SD = 6), all reporting no history of speech, hearing, language, reading, or neurological disorders.

Power. Our pilot study (see Supplement S1 for details) yielded a moderate effect size for the effect of lexicality (d=.55), indicating that a sample size of N = 28 would be sufficient to detect this effect with 80% power. To allow detection of smaller effects ($d \geq .45$) with adequate power, we planned for $N \geq 40$. Finally, to accommodate a 30% exclusion rate, we targeted 58 participants. We note, however, that detecting medium-sized interactions in a 2 × 2 design (f=.25, approximately d≈.5) would require substantially larger samples (≈180 participants in total) to achieve 80% power. Thus, while our study was well-powered to detect the predicted main effect of lexicality, the interaction analyses should be interpreted with caution.

2.3. Overview

The experiment was conducted on participants' computers (mobile devices were not permitted) using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Before the main tasks, participants completed a headphone screening (Woods et al., 2017); those who failed were excluded from further participation. The study comprised two sessions of a VAS task, separated by two language assessments; this took approximately 45 minutes per subject.

2.4. Visual Analog Scaling task

Stimuli. In its entirety, the VAS task included 10 matched sets of word and nonword continua: two bilabial stop voicing pairs, three alveolar stop voicing pairs, two velar stop voicing pairs, and three fricative place pairs (Table 1). Each continuum was constructed from seven acoustic steps, including the minimal pair endpoints. Because articulation may differ in words and nonwords (Scarborough, 2012; Stephenson, 2004), and in words with minimal pairs than words without competitors (Wedel et al., 2018), we also introduced a ***splicing*** manipulation that

led to four continua for each pair, crossing whether the onset portion came from a word or nonword, and whether the stimulus as a whole was a word or nonword.

Each endpoint token was recorded within the carrier sentence "*He said* ____", with five repetitions by a male speaker[1]. Noise reduction was applied to the audio files in Audacity before the target items were segmented from the carrier sentence, and then intensities were scaled to 70 dB across items.

Table 1. List of items used in the VAS task

Word	Nonword
bond-pond	*bonf-ponf*
batch-patch	*bazg-pazg*
tent-dent	*tenf-denf*
tusk-dusk	*tups-dups*
tense-dense	*tench-dench*
cage-gage	*caish-gaish*
coat-goat	*coadge-goadge*
sift-shift	*sigged-shigged*
seep-sheep	*seeg-sheeg*
sip-ship	*sib-shib*

For ***stop voicing continua***, we used a Praat script (Winn, 2020) to generate seven-step VOT continua (in which f0 was selected to covary across the seven steps). The range of VOTs for each stop voicing continuum differed as a function of the place of articulation (following Lisker & Abramson, 1964) with 0-40 msec for /b/-/p/; 5-50 msec for /t/-/d/; 10-60 msec for /k/-/g/.

Fricative place continua were created using a spectral averaging method (Colby et al., 2023; Galle et al., 2019; McMurray et al., 2018). This started by excising the frication portions of /s/–/ʃ/ endpoint recordings, estimating their spectrum, shifting the spectra in frequency space, and then filtering noise through that spectrum. To ensure that coarticulation also varied with continuum step, we cut the vocoid from the fricative and TANDEM-STRAIGHT to shift the formants of the vocoid in 7 even steps (Kawahara et al., 1999) before splicing it onto the frication portion. More detailed stimulus manipulation procedures are provided in Supplemental S2.

Lastly, to ensure that the acoustic cues in the continua were the same across lexical conditions, items were reconstructed by splicing the onset consonant and vowel from each step of the continua onto the coda consonants to create four continua for each set. For each, the onset consonant and vowel from one token (e.g., *ba* from *batch*) was combined with the coda from the same item (e.g., *tch* from *batch;* the *match-splice conditi*on), or the other item (e.g., *zg* from *bazg*; the *cross-spliced condition*). In the match-splice condition, the coda was taken from another recording of the same item. This led to two splice conditions (match-spliced v. cross-spliced) crossed with the word/nonword conditions to ensure that neither articulation nor splicing was a confounding variable.

Procedure. Each VAS trial displayed a horizontal line with orthographic labels of the two endpoints at each extreme. Participants heard an auditory stimulus over headphones/earphones and clicked a point on the line indicating how closely the stimulus matched either endpoint. A vertical tick mark appeared at their selection point. Listeners could revise their responses before continuing. Trials advanced automatically 300 msec after the final response.

Each participant was randomly assigned one continuum per contrast type (bilabial, alveolar, velar stop voicing, fricative place), and completed 448 trials (7 steps × 4 continua × 2 lexicality types × 2 repetitions × 2 splicing conditions × 2 sessions). Trials were blocked by continuum, with words, their corresponding nonwords, and both splice conditions appearing

[1] The talker was a monolingual speaker of American English from the Midwest.

within the same block (e.g., *batch–patch* trials were interleaved with *bazg–pazg* trials). Trial order within blocks and block order were randomized. The assignment of labels to the side of the continua (e.g., *batch* on the left or right) was reversed in the second session. Each session took approximately 15 minutes.

***Data Processing*.** Responses from the second session were reverse-scored so that ratings of 0 and 100 each referred to the same phonemes across sessions. Next, we visually inspected plots of each subject's data (grouped by lexicality and splicing condition); 10 participants were excluded for having flat functions, which could be indicative of lack of engagement in the task.

For the remaining data (48 participants), four-parameter logistic functions were fit to each subject, each continuum, and each splicing condition using a custom curve-fitting program developed in MATLAB (version 48, McMurray, 2017) (https://osf.io/4atgv/). Fits minimized least-squares error with constraints (e.g., crossover within the continuum range, bounded asymptotes) and provided estimates for minimum and maximum asymptotes, slope, crossover, and root mean squared error (RMSE) of a subject's response ratings. Fits were visually inspected, with no exclusions; average model fit was high (mean R^2 = .78).

From this output, the following three variables were extracted: (1) slope, which inversely reflects ***categorization gradiency*** (i.e., a steeper slope reflects less gradiency); (2) response variability (RV), measured by trial-by-trial RMSE between individual responses and the predicted psychometric function, inversely reflecting ***categorization consistency*** (Apfelbaum et al., 2022; Kim, Klein-Packard, et al., 2025): lower RV indicates tighter clustering of responses around the estimated function, reflecting more consistent categorization; and (3) ***amplitude***, the difference between the two asymptote values. Amplitude is an additional index, but its theoretical interpretation is less established compared to RV and slope.[2] Accordingly, all results from amplitude analyses are reported separately in Supplemental S3.

2.5. Measuring General Language Ability (Language assessment tasks)

To assess participants' overall language abilities, we administered two in-house assessments: an adapted Token Test and an Agent-Action Test (see Kim et al., 2024). The adapted Token Test evaluated sentence comprehension. On each trial, participants viewed a 6×2 grid containing 12 objects—circles and squares in six colors (white, black, red, green, blue, and yellow). They then heard a directive (e.g., "*Click on the circles between two squares"*) and responded by clicking or moving shapes. The task included 30 trials of increasing complexity and lasted ~5 minutes.

The Agent-Action Test assessed receptive vocabulary. On each trial, participants heard a true/false question containing low-frequency lexical items (e.g., *Is an esoteric topic widely understood?)* and clicked on "Yes" or "No". The task included 25 randomly ordered trials and lasted fewer than 5 minutes.

The auditory stimuli for both tasks were generated by an AI text-to-speech tool (https://ttsmaker.com/). A full list of test items is available on the OSF page.

Data Processing. Responses for each language test were scored as correct or incorrect (1 or 0). Scores from the two assessments were significantly correlated, $r(46) = .76$, $p < .0001$; thus, we computed a composite by z-scoring the accuracy of each task and averaging them.

---

[2] There remain open questions regarding whether amplitude functions as a primary index of categorization or whether it largely reflects variance already captured by RV, in which case it may be redundant.

3. Results

All analyses were performed in R (R Core Team, 2025) and R Studio (Posit team, 2025). Linear mixed-effects models were implemented using the *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) packages.

3.1. Effect of Lexicality on VAS Indices

The effects of lexicality and splicing condition in each analysis are illustrated in **Figure 1**. Linear mixed-effects models assessed the effect of lexicality on (1) categorization consistency and (2) categorization gradiency. On an exploratory basis, we also examined amplitude (described in Supplement S3). These were z-scored to permit a comparison of the magnitude of the fixed effects across models. The two independent variables were lexicality (sum-coded as word = +1 and nonword = −1) and splice condition (match vs. cross, +1 vs. −1, respectively). Per our pre-registered analysis plan, we started with a model with a maximal random effects structure and simplified it until the model converged (see the *Note* under Table 2 for the model syntax).[3]

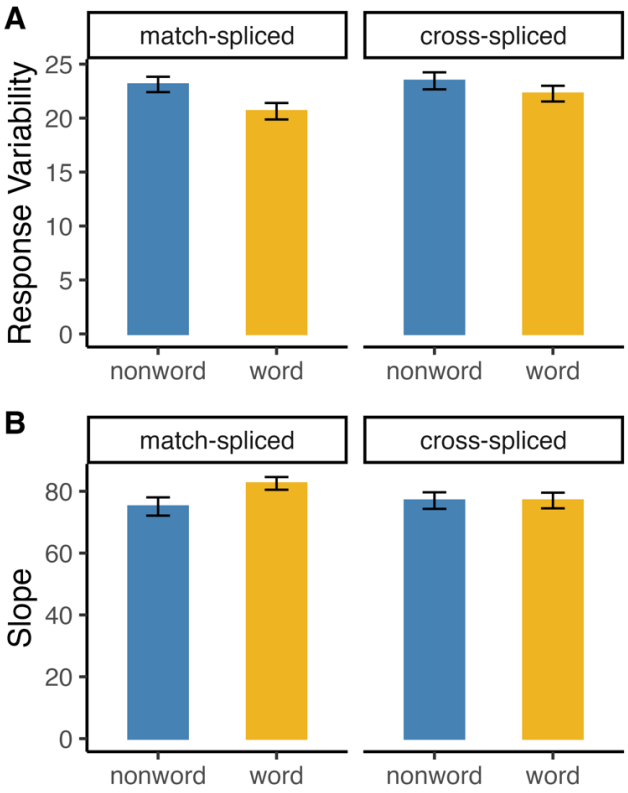

Figure 1. Raw values of VAS parameter estimates as a function of lexicality (x-axis) and splice condition. Each panel displays mean values (±1 SE) for (A) response variability (RV; higher raw scores reflect lower categorization consistency), and (B) slope (higher raw scores reflect more categorical responses).

We started by analyzing ***RV.*** Word trials had significantly lower RV than nonword trials (β = −0.092, $t(621)$ = −3.644, $p < .001$) (Table 2A). This is consistent with our hypothesis that the availability of top-down knowledge (here, lexical knowledge) leads to more consistent categorization. There was no effect of splicing condition (p = .055), suggesting that categorization consistency did not vary significantly as a function of the coarticulatory cues in the productions of these tokens. Finally, the interaction between lexicality and splicing condition was not significant (p = .169), indicating that coarticulation did not modulate the effect of lexicality.

Next, we examined ***slope*** using the same model as above. None of the terms was significant (all ps > .05; Table 2B), suggesting that neither top-down knowledge nor coarticulatory cues affects the extent to which categorization is either categorical or gradient.

[3] Random intercepts were included for subject and continuum. *Continuum*, in this case, is defined as word and nonword pairs for a given contrast; e.g., *bonf-ponf* and *bond-pond* are modeled as a single continuum.

3.2. Individual Differences

Next, we asked how individual differences in language ability related to the effects reported above. Before the main analysis, we conducted a simple regression predicting language scores from RV. This model revealed that categorization consistency alone accounted for 31% of the variance in language ability (R^2 = .31, p < .001), closely mirroring the effect size reported previously.

For the main analysis, we used a linear mixed effects model to predict RV (z-scored) from lexicality and splice condition (sum-coded as before), the composite language score (z-scored), and all two-way interactions. **Figure 2** shows the large correlation between categorization consistency and language scores. Model summaries and model syntax are reported in **Table 3**.

Once again, the results showed that words led to greater categorization consistency (lower RV) than nonwords (β = −0.097, t(621) = −3.631, p < .001). The effect of splicing was not significant (p = .085). The effect of language was such that individuals with higher language scores had higher categorization consistency (β = −0.660, t(46) = 7.452, p < .001). However, there were no significant interactions (all ps > .05): the effect of language on categorization consistency was not modulated by whether it was measured on word or nonword trials.

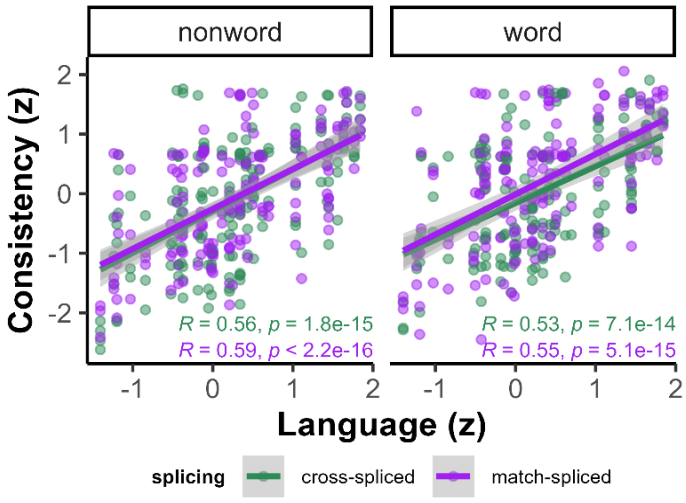

Figure 2. Correlation between categorization consistency and language score as a function of lexicality and splicing condition.

Table 2. Summary of coefficients in the regression models on VAS indices, with lexicality, splicing condition, and their interaction as fixed effects.

A. Response Variability				
Fixed effects	**Estimate**	**SE**	***t***	***p***
(Intercept)	0.005	0.124	0.039	.970
Lexicality	−0.092	0.025	−3.644	< .001
Splicing Condition	−0.049	0.025	−1.926	.055
Lexicality × Splicing Condition	−0.035	0.025	−1.377	.169
B. Slope				
(Intercept)	0.007	0.052	0.141	.890
Lexicality	0.056	0.037	1.507	.132
Splicing Condition	0.027	0.037	0.735	.462
Lexicality × Splicing Condition	0.055	0.037	1.471	.142

Note. R syntax: VAS index.Z ~ Lexicality + SpliceCond + LexType:SpliceCond + (1 | Subject) + (1 | Continuum).

Next, we repeated the analysis with slope as the DV. Again, none of the terms was statistically significant, meaning that the categorical/gradient nature of categorization is not related to language ability, and this did not differ by lexicality.

Table 3. Summary of output from regression models of VAS indices with lexicality, splicing condition, and z-scored language composite scores as fixed effects.

A. Response Variability				
Fixed Effects	**Estimate**	**SE**	***t***	***p***
(Intercept)	0.183	0.097	1.881	.069
Lexicality (word)	−0.097	0.027	−3.631	< .001
Splice Condition (match -spliced)	−0.046	0.027	−1.727	.085
Language Composite	−0.660	0.089	−7.452	< .001
Lexicality x Splice Condition	−0.035	0.024	−1.370	.171
Lexicality x Language	0.017	0.030	0.575	.565
Splice Condition x Language	−0.009	0.030	−0.310	.756
B. Slope				
(Intercept)	−0.002	0.056	−0.042	.967
Lexicality (word)	0.060	0.039	1.527	.127
Splice Condition (match-spliced)	0.016	0.039	0.416	.677
Language Composite	0.037	0.059	0.622	.537
Lexicality x Splice Condition	0.055	0.037	1.467	.143
Lexicality x Language	−0.014	0.045	−0.312	.755
Splice Condition x Language	0.041	0.045	0.923	.356

Note. R syntax: VAS.index ~ Lexicality + SpliceCond + z_LangComp + LexType:SpliceCond + LexType:z_LangComp + SpliceCond:z_LangComp + (1 | Subject) + (1 | Continuum)

4. Discussion

Mounting research with the Visual Analog Scale (VAS) task has revealed that categorization consistency is a critical predictor of real-world language and reading outcomes (Honda et al., 2024; Kim, Klein-Packard, et al., 2025; Kim et al., 2024; Myers et al., 2024). This construct reflects how much a listener's percept rating varies from trial to trial and is not well described by theoretical models of speech perception. To better understand the locus of individual differences in categorization consistency, the current study investigated the effect of higher-level knowledge on categorization consistency. Specifically, we used words and nonwords on separate trials to test the auditory/phonological clean-up (APCU) hypothesis: that higher-level knowledge cleans up lower-level noise and thereby increases categorization consistency.

Consistent with our prediction, categorization consistency was significantly higher on word trials than on nonword trials. Importantly, the splicing manipulation ensured that the segment /b/ in *bonf* on a cross-spliced trial was the same /b/ in *bond* on a match-spliced trial. The results, however, suggest that it was the effect of the coda—whether it formed a word or a nonword with the given onset—that affected how consistently the segment was perceived. Splicing condition (match vs. mismatch) was included in the model, but it neither had a significant main effect nor significantly interacted with lexicality. These findings together provide strong support for the hypothesis that, in the face of bottom-up noise, perception of a segment such as /b/ is stabilized by higher-level knowledge (lexicality, in this case).

Though we propose that the APCU mechanisms function via feedback, this remains speculative. The same observation could arise from clean-up at the level of the lexicon or the mapping between auditory and lexical representations and the VAS response (for an analogical

model, see Norris et al., 2000). Our work serves as a first step: if there had been no difference as a function of lexicality, the APCU hypothesis would be unsupported. While it is premature to definitively say at which level the effect arises, the robust associations in the literature between categorization consistency and language/reading skills is important: it is unclear why differences solely in the response system would explain so much variance in language. Future work should build on this first step with more sophisticated designs and measures that can assess the perceptual representation more directly.

In addition to categorization consistency, we analyzed the more traditional categorization index: slope (gradiency). Slope was not significantly affected by lexicality, splicing condition, or their interaction, implying that whether a listener is gradient or categorical is not affected by higher-level knowledge or coarticulatory cues. This is perhaps not surprising, given that slope is also less robustly correlated across continua: a steep slope for stop voicing categorization does not predict a steeper slope for fricatives (Kim, McMurray, et al., 2025). The present data are consistent with this: slope showed no reliable cross-contrast correlations (words: $r(46) = -.11$, $p = .46$; nonwords: $r(46) = .23$, $p = .112$), whereas categorization consistency was strongly correlated across contrast types (words: $r(46) = .71$, $p < .0001$; nonwords: $r(46) = .75$, $p < .0001$). This suggests that slope may not be a product of larger language function, but rather, derives from differences in a listener's experience with a particular contrast (e.g., Clayards et al., 2008; Kim, McMurray, et al., 2025; Kutlu et al., 2024). However, the lack of an effect on slope also provides some discriminant validity regarding the APCU hypothesis, which predicts an effect specifically on categorization consistency.

A second aim was to evaluate how the effect of lexicality related to language ability. We evaluated language ability by using a composite language score of the same tests used in prior work (Kim et al., 2024). Here, we predicted an interaction with lexicality that would suggest either that the influence of higher-level knowledge on cleaning up lower-level noise could be dependent on greater language ability, or that it could be stronger for individuals with lower language ability as a manifestation of a compensatory mechanism.

We replicated the finding that individuals with higher language scores also show higher categorization consistency (Kim et al., 2024). However, there was no significant interaction with lexicality, providing no support for either of the proposed interpretations. One possibility is that word and nonword categorization reflect a large amount of shared variance. In fact, despite the significant effect of lexicality on RV, RV was strongly correlated between the word and nonword continua, $r(46) = .92$, $p < .0001$. Slope was also significantly, though less strongly, correlated, $r(46) = .52$, $p < .001$. To some extent, such correlations were predicted—the word and nonword continua used the same auditory instantiations, and certainly some variance in slope and RV reflects specific response to specific auditory cues. However, the magnitude of the correlation in RV was unexpected and does not favor a model in which differences in the ability to engage APCU can explain the high correlation between language and categorization consistency. This echoes previous work by Kim, McMurray, et al. (2025), who found that individuals' RV was more strongly correlated across continua than their slopes, suggesting that consistency is perhaps "trait-like".

We see two possible explanations for the lack of a moderation by language. First, although we expected the effect of APCU mechanisms on consistency to be moderated by language ability, at face value, our data are more consistent with the idea that categorization consistency is trait-like within individuals and individual differences in this trait are largely driven by differences in the auditory/perceptual system. APCU then acts relatively uniformly (across individuals) to clean up some of the inherent noise in the system.

An alternative explanation is that the experimental design limited the extent to which individual variation in the APCU effect was apparent. For example, several factors could have encouraged lexicalization of the nonwords over the course of the experiment: their orthographic presentations at the ends of the scale; the use of monosyllabic, phonotactically legal stimuli; and

repeated exposure. Indeed, the nonwords were fairly word-like (e.g., using a common CVCC structure, frequent consonants), thus providing APCU mechanisms a great deal of information to leverage. Follow up work should examine nonwords that are less familiar. Similarly, the use of just one talker likely promoted talker familiarity, a higher-level cue that the APCU mechanisms could leverage even in the context of nonwords.

Broadly, this study demonstrated a clear effect of lexicality on categorization consistency. Though more work is needed, it provides the first piece of support for the hypothesis that categorization consistency reflects, in part, the effect of APCU mechanisms. Future work may seek to clarify the role of feedback in APCU and to identify individual differences in the strength of these mechanisms.

References

Anwyl-Irvine, A. L., Massonnie, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behav Res Methods*, *52*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Apfelbaum, K. S., Kutlu, E., McMurray, B., & Kapnoula, E. C. (2022). Don't force it! Gradient speech categorization calls for continuous categorization tasks. *The Journal of the Acoustical Society of America*, *152*(6), 3728-3745.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P., & Bolker, M. B. (2015). Package 'lme4'. *convergence*, *12*(1), 2.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS biology*, *14*(11), e1002577.

Clayards, M. (2018). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *The Journal of the Acoustical Society of America*, *144*(3), EL172-EL177.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804-809.

Colby, S., Seedorff, M., & McMurray, B. (2023). Audiological and Demographic Factors that Impact the Precision of Speech Categorization in Cochlear Implant Users. *Ear and hearing*, *44*(3), 572-587.

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*, e229.

Fuhrmeister, P., & Myers, E. B. (2021). Structural neural correlates of individual differences in categorical perception. *Brain and Language*, *215*, 104919.

Galle, M. E., Klein-Packard, J., Schreiber, K., & McMurray, B. (2019). What are you waiting for? Real-time integration of cues for fricatives suggests encapsulated auditory memory. *Cognitive science*, *43*(1), e12700.

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

Getz, L. M., & Toscano, J. C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological science*, *30*(6), 830-841.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218-1227.

Honda, C. T., Clayards, M., & Baum, S. R. (2024). Exploring individual differences in native phonetic perception and their link to nonnative phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(4), 370.

Hornickel, J., & Kraus, N. (2013). Unstable representation of sound: a biological marker of dyslexia. *Journal of Neuroscience*, *33*(8), 3500-3504.

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594.

Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-

frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, *27*(3-4), 187-207.
Kim, H., Klein-Packard, J., Sorensen, E., Oleson, J., Tomblin, J. B., & McMurray, B. (2025). Speech Categorization Consistency Predicts Language and Reading Abilities in School-Age Children: Implications for Language and Reading Disorders. *263*(106194). https://doi.org/https://doi.org/10.1016/j.cognition.2025.106194
Kim, H., McMurray, B., Sorensen, E., & Oleson, J. (2025). The consistency of categorization-consistency in speech perception. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-025-02700-x
Kim, H., Tomblin, B., & McMurray, B. (2024). Speech Categorization Consistency Predicts Overall Languae Abilities. https://doi.org/https://doi.org/10.31234/osf.io/u46pj
Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40-57.
Kutlu, E., Baxelbaum, K., Sorensen, E., Oleson, J., & McMurray, B. (2024). Linguistic diversity shapes flexible speech perception in school age children. *Scientific reports*, *14*(1), 28825.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.
Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.
Lupyan, G. (2017). Objective effects of knowledge on visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(4), 794.
Luthra, S., Crinnion, A. M., Saltzman, D., & Magnuson, J. S. (2024). Do They Know It's Christmash? Lexical Knowledge Directly Impacts Speech Perception. *Cognitive science*, *48*(5), e13449.
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. (2025). TRACE-ing fixations in the Visual World Paradigm: Extending linking hypotheses and addressing individual differences by simulating trial-level behavior. *Brain Research*, *1856*, 149563.
Magnuson, J. S., Crinnion, A. M., Luthra, S., Gaston, P., & Grubb, S. (2024). Contra assertions, feedback improves word recognition: How feedback and lateral inhibition sharpen signals over noise. *Cognition*, *242*, 105661.
Magnuson, J. S., & Luthra, S. (2024). Simple recurrent networks are interactive. *Psychonomic Bulletin & Review*, 1-9.
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*, 369.
Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech communication*, *2*(1), 15-35.
McHaney, J. R., Tessmer, R., Roark, C. L., & Chandrasekaran, B. (2021). Working memory relates to individual differences in speech category learning: Insights from computational modeling and pupillometry. *Brain and Language*, *222*, 105010.
McMurray, B. (2017). Nonlinear curvefitting for Psycholinguistics (and other) Data. *Version 48*.
McMurray, B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America*, *152*(6), 3819-3842.
McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental psychology*, *54*(8), 1472.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological review*, *118*(2), 219.
McMurray, B., & Jongman, A. (2016). What comes after/f/? Prediction in speech derives from data-explanatory processes. *Psychological Science*, *27*(1), 43-52.
Munson, B., Schellinger, S. K., & Edwards, J. (2017). Bias in the perception of phonetic detail in children's speech: A comparison of categorical and continuous rating scales. *Clinical Linguistics & Phonetics*, *31*(1), 56-79.
Myers, E., Phillips, M., & Skoe, E. (2024). Individual differences in the perception of phonetic category structure predict speech-in-noise performance. *The Journal of the Acoustical Society of America*, *156*(3), 1707-1719.
Neef, N. E., Schaadt, G., & Friederici, A. D. (2017). Auditory brainstem responses to stop consonants predict literacy. *Clinical Neurophysiology*, *128*(3), 484-494.
Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 873.
Noe, C., & Fischer-Baum, S. (2020). Early lexical influences on sublexical processing in speech perception: Evidence from electrophysiology. *Cognition*, *197*, 104162.
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.
Ou, J., Yu, A. C., & Xiang, M. (2021). Individual differences in categorization gradience as predicted by online processing of phonetic cues during spoken word recognition: Evidence from eye movements. *Cognitive science*, *45*(3), e12948.
Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, *39*(3), 347-370.
Posit team. (2025). *RStudio: Integrated Development Environment for R.* In Posit Software, PBC. http://www.posit.co/
R Core Team, R. (2025). R: A language and environment for statistical computing. In: Citeseer.
Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological science*, *12*(4), 348-351.
Sarrett, M. E., McMurray, B., & Kapnoula, E. C. (2020). Dynamic EEG analysis during language comprehension reveals interactive cascades between perceptual processing and sentential expectations. *Brain and language*, *211*, 104875.
Scarborough, R. (2012). Lexical similarity and speech production: Neighborhoods for nonwords. *Lingua*, *122*(2), 164-176.
Stephenson, L. (2004). Lexical frequency and neighbourhood density effects on vowel production in words and nonwords. Proceedings of the 10th Australian International Conference on Speech Science and Technology,
Stilp, C. (2020). Acoustic context effects in speech perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(1), e1517.
Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language*, *100*, 61-88.
Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *The Journal of the Acoustical Society of America*, *147*(2), 852-866.
Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064-2072.

Lexical Knowledge Enhances Consistency in Speech Categorization

ONLINE SUPPLEMENT

Sita Carraturo [a*], Hyoju Kim [b], Ethan Kutlu [b, c, d], Bob McMurray [b, c, d]

[a] Department of Otolaryngology, University of Iowa, Iowa City, IA 52242, USA
[b] Department of Psychological & Brain Sciences, University of Iowa, Iowa City, IA 52242, USA
[c] Department of Linguistics, University of Iowa, Iowa City, IA 52242, USA
[d] Department of Communication Sciences & Disorders, University of Iowa, Iowa City, IA 52242, USA

* Corresponding author: Sita Carraturo, Department of Psychological & Brain Sciences, University of Iowa, PBSB 255E, Iowa City, IA, 52242, United States of America.

E-mail addresses: sita-carraturo@uiowa.edu (S. Carraturo), hyoju-kim@uiowa.edu (H. Kim), ethan-kutlu@uiowa.edu (E. Kutlu), bob-mcmurray@uiowa.edu (B. McMurray).

S1. Pilot Data

To inform the design of the present study, we conducted two pilot experiments examining whether lexical status (word vs. nonword) modulates phonetic categorization patterns, using a Visual Analog Scaling (VAS) task. Both experiments were administered online via Gorilla Experiment Builder and recruited participants through Prolific.

In **Pilot Experiment 1**, forty-eight adult native speakers of American English (30 female; mean age = 31.6 years, SD = 5.4) completed the task. The stimuli comprised four-word continua (*beach–peach, dime–time, gold–cold, dark–bark*) and four corresponding nonword continua (*beag–peag, dich–tich, golss–kolss, darp–barp*), representing two stop voicing and two stop place contrasts. Each continuum was manipulated into nine acoustic steps. To ensure that lexical status was the only systematic difference between paired conditions, the manipulated portions of the word stimuli were spliced onto their corresponding nonword pairs, rendering each word–nonword pair acoustically identical in the critical region.

The results revealed no significant difference in categorization slope between word and nonword continua, $t(47) = -1.14$, $p = .26$ (Figure S1A). However, categorization consistency was significantly higher for word continua compared to nonwords, $t(47) = -3.77$, $p < .001$ (Figure S1B). Amplitude measures also differed by lexical status, $t(47) = -3.67$, $p < .001$, with greater amplitude observed in responses to word continua (Figure S1C). A linear mixed-effects model predicting each VAS index from lexical status (as a fixed effect), with subject as a random intercept, confirmed a significant effect of lexical status on categorization consistency ($p < .001$) but not on slope. These findings suggest that lexical knowledge enhances the stability of speech categorization. However, this pilot did not include a non-spliced baseline condition, limiting our ability to determine whether the observed effects could be attributed to coarticulatory cues. This limitation motivated a second pilot experiment.

Pilot Experiment 2 addressed this gap by including both spliced and non-spliced versions of the stimuli and by broadening the range of phonological contrasts. Forty-one adult native speakers of American English (20 female; mean age = 29.8 years, SD = 5.3) participated. The stimulus set comprised four-word continua (*beach–peach, save–shave, check–chuck, least–list*) and four matched nonword continua (*beesp–peesp, sague–shague, chesp–chusp, leedge–lidge*), which included one stop voicing contrast, one fricative place contrast, and two vowel contrasts. Each continuum contained seven steps. As in the first pilot, continua were cross-spliced to ensure acoustic equivalence across lexical conditions, and a non-spliced version of each continuum was also included.

The results replicated the main findings of Pilot 1 (Figure S1D-F). A mixed-effects model predicting categorization consistency from lexical status (fixed effect), with subject as a random intercept, revealed a significant effect of lexical status ($\beta = -2.07$, $SE = 0.60$, $p = .001$), again indicating higher consistency for word continua. No significant effect was found for slope. A second model included lexical status, splicing condition (spliced vs. non-spliced), and their interaction. This model revealed only a significant main effect of lexical status, with no effect of splicing condition or interaction. These results indicate that top-down lexical information enhances perceptual stability irrespective of the presence of coarticulatory cues.

Despite these promising findings, the second pilot remained limited by the small number of contrast pairs and the inclusion of vowel contrasts, which proved challenging for acoustic manipulation. Vowels are highly coarticulated segments, making it difficult to create natural-sounding, tightly controlled stimuli.

Together, the results from both pilot experiments provide converging evidence that lexical knowledge stabilizes categorization consistency, independent of coarticulatory context. These findings informed the design of the current study by motivating the use of a new and more extensive set of continua focused on stop voicing and fricative place contrasts—dimensions that allow for more precise acoustic manipulation and theoretical clarity.

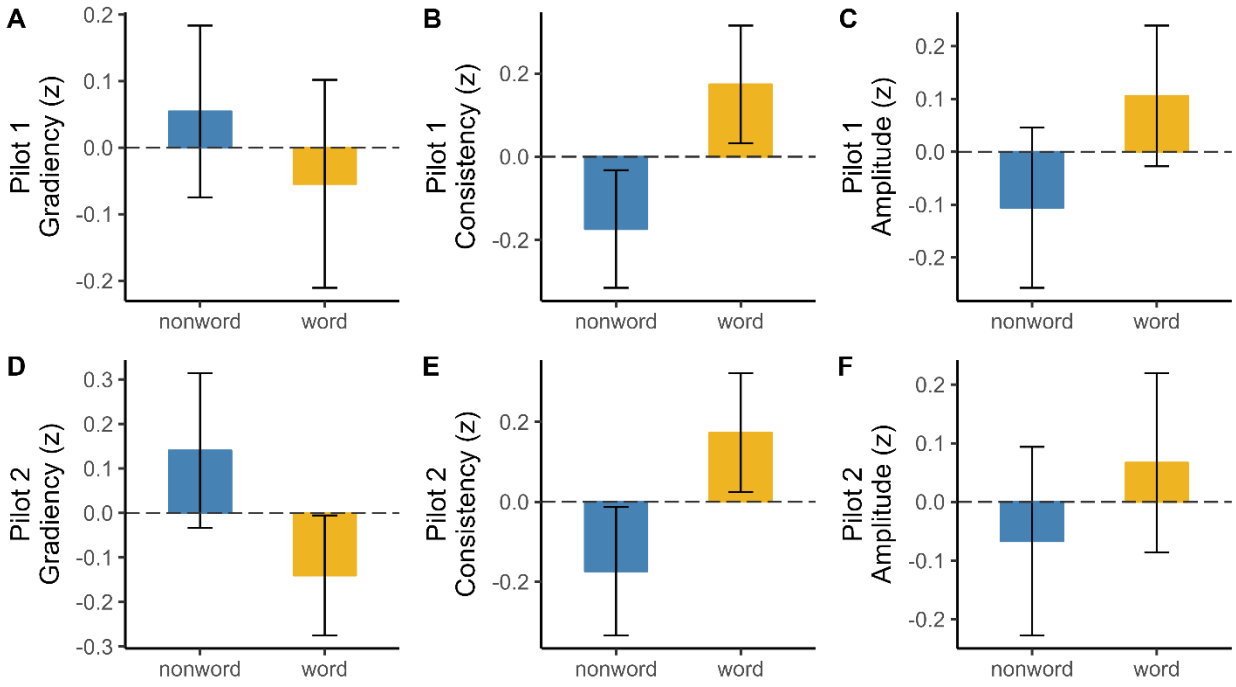

Figure S1. Z-scored VAS parameter estimates as a function of lexical type (blue = nonword; yellow = word). Each panel displays mean values (±1 SE) for three VAS-derived measures. The first three panels show the results from the first pilot study: (A) gradiency (inverse slope), (B) categorization consistency (inverse RV), and (C) amplitude. The last three panels show the results from the second pilot study: (D) gradiency (inverse slope), (E) categorization consistency (inverse RV), and (F) amplitude.

S2. Additional Stimulus Creation Details

For stop voicing continua, two recordings of each word were selected. The initial 100 msec of silence, the initial consonant, and the vowel from one file of each end of a continuum (e.g., *ba* from one recording of *batch* and the *pa* from one recording of *patch*) were isolated and run through a Praat script (Winn, 2020) to create a seven-step VOT continuum. Each step of this *ba/pa* continuum was then spliced together with the coda and final 100 msec of silence from the other recording of the same word (e.g., *tch* from the other recording of *batch*), generating a "match-spliced" continuum. Then, the original seven steps were spliced together with the coda and 100ms of silence from the corresponding nonword (e.g., *zg* from *bazg*) to create a "cross-spliced" continuum; the same is done for nonwords (e.g., after making a 7-step continuum from isolated *ba/pa* splices from *bazg/pazg* recordings, each step is spliced with *zg* from another recording of *bazg* and cross-spliced with *tch* from *batch*).

For fricative place continua, spliced endpoints were created following the same procedure as described above. To create /s/-/ʃ/ continua, we used a spectral averaging procedure developed in prior studies (Colby et al., 2023; Galle et al., 2019; McMurray et al., 2018). First, the frication portions from endpoint tokens (e.g., save and shave) were extracted from the selected recordings. Second, the longer frication segment was cut to match the length of the shorter one, ensuring consistency in segment length. Third, the spectral mean is calculated from the long-term average spectra of each fricative, and both spectra were aligned to the same average spectral mean. Fourth, a weighted average of the two spectra is constructed to create a series of seven spectra, representing each step along the continuum. Fifth, the spectral means of the spectra are shifted in frequency space to create seven steps. Sixth, the modified spectra are applied as filters to a segment of white noise, which has an envelope that is the average of the /s/ and /ʃ/ endpoints. These steps were implemented in MATLAB script (https://osf.io/ut9wz/). Separately, a continuum of the vocoids of the endpoints was created in TANDEM-STRAIGHT (Kawahara et al., 1999). Lastly, we splice each step of this continuum onto the corresponding /s/-/ʃ/ continuum. Cross-spliced continua were created by removing the coda from each step of the continua and splicing the coda of the non-word onto the word and vice-versa (e.g., the /p/ from *sheep* onto *sheeg* and the /g/ from *sheeg* onto *sheep*).

S3. Amplitude Analyses and Results

First, we examined the effects of lexical type and splicing on amplitude (Table S3A). Word continua, on average, had significantly higher amplitudes than nonword continua (β = 0.115, t(621) = 4.365, p < .001). Similarly, match-spliced trials showed significantly higher amplitudes than cross-spliced trials (β = 0.077, t(621) = -2.931, p < .01). This suggests that coarticulatory cues consistent with a natural production of a token lead to more robust categorization of endpoint tokens (or conversely that mismatching coarticulation can disrupt the ambiguous tokens). The interaction between lexical type and splicing condition was not significant (p = 0.190).

Next, we analyzed how language scores modulated the relationships of lexical type and splicing on amplitudes (Table S3B). The analysis showed that both lexical type and splicing conditions affected amplitude. Words again showed larger amplitudes than nonwords (β = .135, t(619) = 4.905, p < .001) and match-spliced tokens had larger amplitudes than cross-spliced ones (β = 0.080, t(619) = 2.926, p < .01).

There was also a main effect of language ability such that those with higher language scores also yielded more robust categorization of endpoint tokens (β = 0.357, t(46) = 3.273, p < .01). Finally, there was also a significant interaction between language ability and lexical type such that the effect of lexical type on amplitude was attenuated for people with higher language scores (β = −0.076, t(619) = −2.429, p < .05). **Figure S3** shows that this interaction is characterized by a sort of ceiling effect among individuals with high language scores, for whom amplitudes are already high, independent of lexical type.

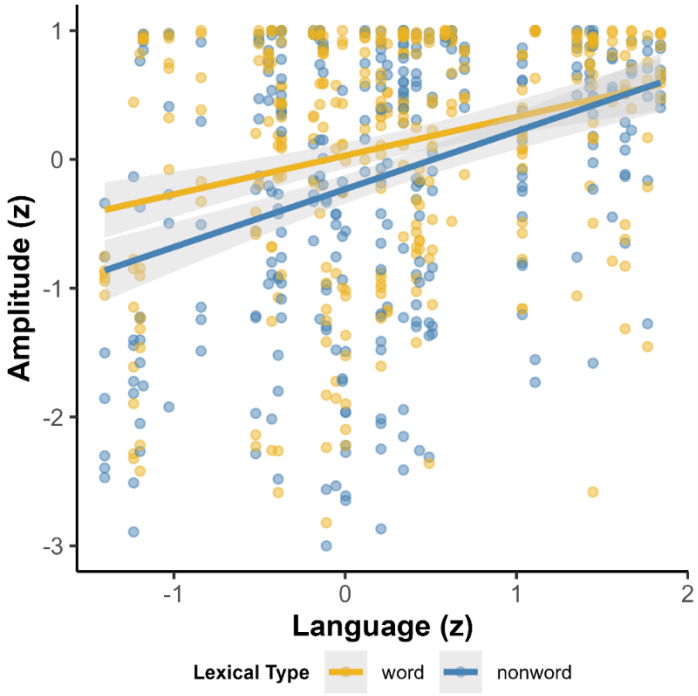

Figure S3. Interaction effect of lexical type and language score on amplitude. Lines show linear regression with 95% confidence interval bands in gray.

In summary, lexical type had a significant effect on amplitude, which was higher in the word condition. That lexical type was unrelated to slope but related to both amplitude and consistency may indicate inherent relationships between these measures: whereas differences in slope are *further characterized* by consistency (Apfelbaum et al., 2022), high consistency may be a prerequisite for high amplitudes, but this remains to be investigated empirically.

Table S3. Summary of output from regression models of amplitude.

A. Amplitude ~ LexType + Splicing + LexType:Splicing				
(Intercept)	0.024	0.132	0.180	.858
Lexical type	0.115	0.026	4.365	< .001
Splicing Condition	0.077	0.026	2.933	< .01
Lexical type × Splicing Condition	0.035	0.026	1.343	.190
B. Amplitude ~ LexType + Splicing + Lang + LexType:Splicing + Lang:LexType + Lang:SpliceCond				
(Intercept)	−0.072	0.128	−0.566	.576
Lexical Type (word)	0.135	0.027	4.905	< .001
Splice Condition (match-spliced)	0.080	0.027	2.926	< .01
Language Composite	0.357	0.109	3.273	< .01
Lexical Type x Splice Condition	0.036	0.026	1.361	.174
Lexical Type x Language	−0.076	0.031	−2.429	< .05
Splice Condition x Language	−0.013	0.031	−0.424	.672