Journal of Memory and Language 103 (2018) 58–73

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Listener sensitivity to probabilistic conditioning of sociolinguistic variables: The case of (ING)

Charlotte Vaughn*, Tyler Kendall

University of Oregon, Department of Linguistics, 1290 University of Oregon, Eugene, OR 97403-1290, USA

ARTICLE INFO

Keywords:
Speech perception
Sociolinguistic variables
Expectation
Listener knowledge
Grammatical conditioning
Stylistic/phonetic congruence

ABSTRACT

This paper investigates the extent to which listeners are cued into the systematicity of variability in speech, particularly the grammatical conditioning constraints of the English sociolinguistic variable (ING) (e.g., *talking* vs. *talkin*). Listeners' sensitivity to the realization of (ING) words embedded in sentences was tested under various conditions. Comprehenders demonstrated expectations about the grammatical category constraints conditioning the realization of (ING) even though such knowledge may not be very informative about word recognition (Experiment 1). As more reliable phonetic information was available, listeners weighted grammatical expectations less, favoring other cues in the signal (Experiments 2 and 3). Bridging a gap between psycholinguistic and sociolinguistic accounts of probabilistic conditioning, these results suggest that factors beyond utility for word recognition may contribute to the probabilistic monitoring of a variable, and underscore the opportunistic nature of the speech comprehension system, where listeners make use of whatever information they have to process the signal.

Introduction

During speech perception, listeners regularly encounter variability across many acoustic cues. The process of language comprehension more generally is thought to be guided by comprehenders' prior experiences, which provide information about what interpretations of the signal are most likely (cf. Fine, Jaeger, Farmer, & Qian, 2013; Zarcone, Van Schijndel, Vogels, & Demberg, 2016). For example, listeners track probabilistic associations between the linguistic and social contexts in which sounds occur, and how sounds are realized (e.g., Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2006). Listeners can then use these probabilistic expectations to aid in the task of ambiguity resolution, a central component of language comprehension (cf. Kuperberg & Jaeger, 2016). However, phonetic variability in production does not *always* result in mismatches that could result in ambiguous interpretations or disrupt word recognition; for many variables studied by sociolinguists, such as the stopping of interdental fricatives (e.g., *dat* for *that*) and velar nasal fronting, also known as variable (ING) (e.g., *runnin* for *running*), listeners may not need to accurately interpret the phonetic realization of the variable in order to accurately recognize words. Thus, although it is clear that listeners' expectations based on probabilistic conditioning are valuable for resolving systematic variation that can result in misparsing (e.g., a gross category mismatch along the VOT continuum), it is unknown whether listeners form and use expectations about the kind of systematic variation that does *not* create ambiguity as a rule (e.g., encountering *–in* for *–ing*). In this study, we explore an example of the latter type of systematic variation, variable (ING), to ask whether listeners still form and use expectations about what conditions variation.

The sociolinguistic research tradition has established that (ING) realization depends probabilistically on both social (e.g., gender) and linguistic (e.g., grammatical category) conditioning factors (e.g., Campbell-Kibler, 2007; Fischer, 1958; Forrest, 2015; Hazen, 2008; Houston, 1985; Kendall, 2010, 2013; Labov, 1966, 2001; Tagliamonte, 2004; Wald & Shopen, 1981). In particular, grammatical category conditioning constraints have been shown to be the most important linguistic predictor of (ING) realization across communities, such that progressive verb forms (e.g., "he's *lying*") are more likely to be realized as *–in* than, for instance, adjectival or nominal forms (e.g., "an unexpected *ending*", which will be referred to as "noun-like" forms). Gerunds (e.g., "I like *cooking*") are often found to pattern in-between verb-like and noun-like forms (Kendall, 2013; Tagliamonte, 2004). And, certain pronouns, namely *something* and *nothing*, have been shown not to follow expected noun-like patterns, exhibiting high *–in* rates, while *anything* and *everything* strongly disfavor *–in* (Kendall, 2010; Labov, 2001). A range of other linguistic factors have also been examined (e.g.,

* Corresponding author.
E-mail addresses: cvaughn@uoregon.edu (C. Vaughn), tsk@uoregon.edu (T. Kendall).

https://doi.org/10.1016/j.jml.2018.07.006
Received 1 August 2017; Received in revised form 22 July 2018

phonological environment, number of syllables and word frequency of (ING) word) but none have been as consistently found to influence (ING) realizations as strongly as grammatical category, and we concentrate on (ING)'s grammatical conditioning constraints in this study.

Despite the extensive evidence for grammatical category effects on (ING) in production, it is not known whether these probabilistic conditioning factors are also used by listeners in perception, though the possibility of such a relationship has long been suspected (cf. Bresnan & Ford, 2010). Thus, we investigate whether probabilistic patterns of (ING) variation in production are detectable in comprehension. In doing so, the present study addresses two main research questions: (RQ1) How sensitive are comprehenders to probabilistic constraints conditioning (ING) realization?, and (RQ2) When are comprehenders more or less likely to make use of such information?

RQ1: How sensitive are comprehenders to probabilistic constraints conditioning (ING) realization?

Comprehenders have been shown to use the systematic variability inherent in speech to generate probabilistic expectations about upcoming information, which facilitates language comprehension (see Fine et al., 2013; Zarcone et al., 2016 for recent reviews). Comprehenders' use of expectations has been demonstrated on a variety of levels of linguistic processing, from syntactic (e.g., Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell, Tanenhaus, & Kello, 1993) to phonetic (Allen & Miller, 2004; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Theodore, Myers, & Lomibao, 2015) to social (Van Berkum, Van den Brink, Tesink, Kos, & Hagoort, 2008; Walker & Hay, 2011). For example, comprehenders are less likely to succumb to a garden path effect if the main verb is most often used in an embedded-clause context (which would be congruent with the correct reading of the sentence) than if the main verb is most often used in direct object contexts (which would point down the garden path) (cf. Garnsey et al., 1997; Trueswell et al., 1993). Knowledge about probabilistic conditioning may reduce the search space such that not all candidate lexical items need to be given equal consideration as potential matches.

The work reviewed in the previous paragraph generally investigates variability that regularly causes potential problems for the comprehender, lexical ambiguity or gross category mismatches. However, as mentioned above, the alternation between *–ing* and *–in* is unlikely to have major consequences for word recognition for several reasons. First, both *–ing* and *–in* realizations are mapped to the same referent (i.e., *running* and *runnin* have the same propositional meaning). And, (the few) minimal pairs resulting from the *–ing/–in* alternation are unlikely to cause confusion, especially in sentential context (e.g., *bobbing/bobbin* or *robbing/robin*). Second, since spoken words are processed incrementally, information presented at word onset is known to constrain lexical recognition more than information later in the word (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Marslen-Wilson & Zwitserlood, 1989). Therefore, when hearing an (ING) word, the listener is likely to settle on the accurate resolution of the word before actually reaching the (ING) variant (though cf. Luce, 1986). Finally, (ING) variants represent a nasal place distinction, which is known to be hard to perceive (Hura, Lindblom, & Diehl, 1992), likely especially so in running speech. This collection of factors, then, makes (ING) a prime case of a variable whose variants are largely irrelevant for accurate word recognition.

How would such a variable be treated in speech perception? Many current models of prediction and expectation in speech perception suggest that the primary factor relevant in weighting the utility of tracking and then forming expectations about a variable is its importance for word recognition (e.g., Baayen, Hendrix, & Ramscar, 2013; Cutler, 2012; Goldinger & Azuma, 2004; Kleinschmidt & Jaeger, 2015; Levy, 2008; McMurray & Jongman, 2011). Certainly, assessing linguistic variants' utility for linguistic processing is of primary interest to language comprehension. However, given an assumption of finite cognitive resources, theories must predict the conditions under which the system tracks and uses conditional probabilities. In this paper we ask: to what end this is done? This question is less about how specific models of prediction could account for a potential finding that listeners are sensitive to the grammatical category conditioning of (ING), but rather whether models need to account for such a result in the first place. If comprehenders do show evidence of sensitivity to (ING)'s constraints, even though that information is not particularly important for lexical access, this indicates either that much more probabilistic information is stored than is used for word recognition, or that theoretical models of prediction and expectation ought to entertain that comprehenders' behavior is affected by factors beyond just what is useful for word recognition.

Work on variables similar to (ING) that do not generally create ambiguity suggests that listeners may indeed generate expectations about the realization of (ING) based on probabilistic linguistic conditioning. Listeners are sensitive to context-specific variant frequencies of allophonic variation in variables like flapping and schwa deletion (e.g., Bürki, Ernestus, & Frauenfelder, 2010; Connine, 2004; Connine, Ranbom, & Patterson, 2008; Pinnow & Connine, 2014; Pitt, 2009; Pitt, Dilley, & Tat, 2011). Connine et al. (2008), for instance, found that words with high schwa-deletion rates in a corpus showed different patterns of processing than words with low schwa-deletion rates, reflecting listener sensitivity to lexically-specific phonological conditioning of variants.

Further, we know that listeners socially evaluate speakers based on how they use variation (e.g., Bresnahan, Ohashi, Nebashi, Liu, & Shearman, 2002; Preston, 2013), and there is suggestive evidence that this is true for how variation is conditioned as well (Bender, 2005). In terms of our variable of interest, previous findings show that listeners are sensitive to rates of (ING) realization in their assignment of social judgments to speakers (e.g., Campbell-Kibler, 2007, Labov et al., 2011; Levon & Fox, 2014; Loudermilk, 2013). For example, using a matched guise paradigm, Labov et al. (2011) found that as a speaker's *–in* rate increased, listeners judged them as less suitable to be a TV newscaster. Thus, although evidence indicates that listeners are capable of attending to differences in rates of (ING) realization when asked to make social judgments, it is not known whether listeners' percepts of (ING) realizations are conditioned on the probabilistic production norms of (ING) realization, or whether these conditioning factors are used in making linguistic judgments.

Based on the work reviewed in this section, it may be that the kind and amount of probabilistic conditioning information used to generate expectations that guide processing is not *solely* determined by how informative the information is for word recognition. Thus, for RQ1 we predict that listeners will demonstrate expectations about the grammatical factors that condition when a given (ING) word will be realized as *–in* versus *–ing*.

RQ2: When are comprehenders likely to use expectations about probabilistic linguistic conditioning?

Any finding that comprehenders' expectations about (ING) realization are in line with constraints in production would support our hypothesis for RQ1. But, even if that is confirmed, it is likely that comprehenders' behavior will not *always* reflect what they have tracked. Thus, RQ2 asks when comprehenders will be more or less likely to make use of their top-down information about (ING) variation.

Previous work suggests that comprehenders are differentially reliant on their previous experience-based expectations (equivalent here to top-down information) depending on various factors. That is, first, prediction based on top-down information is most likely to occur when uncertainty about the bottom-up information is highest (e.g., Brouwer, Mitterer, & Huettig, 2013; Connine & Clifton, 1987; Connine et al., 2008; Ganong, 1980; Huettig, 2015; Kuperberg & Jaeger, 2016; McQueen & Huettig, 2012; Pitt & Samuel, 1993; Warren, 1970). For example, McQueen & Huettig (2012) found that comprehenders downweight reliance on a particular cue, even one that is valuable in the current stimulus, when they are in a context where that cue is less reliable overall. Second, different types of speech situations or attentional goals may induce different types of processing (e.g., Davis & Johnsrude, 2007; McAuliffe & Babel, 2016; Sumner, 2013; Sumner et al., 2013; Theodore, Blumstein, & Luthra, 2015; Wild et al., 2012). For example, Sumner and colleagues (Sumner et al., 2013; Sumner, 2013) suggest that casual speech styles may induce more top-down processing, while careful speech styles may induce more bottom-up processing.

This previous work makes key predictions about our RQ2: If listeners do show sensitivity to probabilistic conditioning constraints ("top-down" information), they should show more sensitivity in cases where there is less certainty about the acoustic realization of (ING). To test the extent to which comprehenders rely on grammatical category knowledge, we alter the availability and congruence of acoustic cues across our experiments (Experiment 1 providing no bottom-up acoustic cues to (ING) realization, Experiment 3 providing complete reliability of acoustic cues, and Experiment 2 falling in-between).

Experiment 1: Text

Experiment 1 asked whether participants generate expectations about how the (ING) variable is likely to be realized when given a written sentential frame without any auditory information. We examine whether production conditioning factors predict participants' ratings of how surprised they would be to hear *–in* in various contexts. Participants' intuitions about (ING) realization, captured in Likert-scale ratings, are taken as a proxy for their latent knowledge of conditioning factors in production. While our ultimate interest is listeners' expectations about spoken language, this text-based experiment assesses listeners' baseline knowledge when entering into those situations.

Our main question for Experiment 1 is whether comprehenders show patterns of expectations that are consistent with grammatical category constraints in production. Based on production patterns, we expect that there will be an overall *–ing* bias: *–ing* is always an available option for the realization of (ING), whereas *–in* is more marked and is expected to be more surprising in all grammatical categories. Thus, we discuss the strength of the *–ing* expectation, rather than grammatical categories that favor *–ing* versus categories favoring *–in*. Based on production norms described above, noun-like grammatical categories (e.g., adjectives, nouns, and some pronouns) should generate *strong* –ing *expectations*, while verb-like categories should generate *less strong* –ing *expectations* (progressives) with gerunds somewhere in-between, and the pronouns *something* and *nothing* not patterning with other noun-like forms.

Method

Participants

Ninety-seven participants were recruited using Amazon's Mechanical Turk, and were paid for their time. Using Mechanical Turk's standard filters, the survey was limited to workers located in the U.S. and with Masters status (workers certified by Amazon as demonstrating high-quality work), who had more than 100 HITs approved. Participants were required to be native speakers of English from the U.S. or Canada. Fourteen participants who completed the task were removed from the analysis for giving categorical responses or otherwise failing to follow directions, leaving a total of 83 participants for analysis. Seven of these 83 participants reported that they did not consider themselves to be monolingual, but all reported English as their first language.

Stimulus design

Stimuli included a total of 96 critical sentences (as well as 10 additional sentences, which were used as practice sentences for later experiments and attention checks, and are not analyzed here in keeping with later experiments), each containing one (ING) word (see Appendix A for full list of stimuli). Sentence stimuli were created to be thematically neutral and relatively syntactically simple, and the position of the (ING) word in the sentence differed across stimuli.

Table 1
Stimulus characteristics.

Grammatical category	Progressive verb (N = 48) Gerund (N = 11) Adjectives (N = 16) Noun (N = 9) Pronoun-2 (N = 6) Pronoun-3 (N = 6)
Preceding environment	Alveolar (N = 41) Velar (N = 16) Other (N = 39)
Following environment	Alveolar (N = 29) Velar (N = 24) Other (N = 43)
(ING) Word length (Syllables)	2 (N = 73) 2.5 (N = 1; *frightening*) 3 (N = 16) 3.5 (N = 3; *interesting*) 4 (N = 4)

Stimuli varied along a range of conditioning factors, with particular attention paid to grammatical category. Although our major aim is to determine whether the pattern of noun-like and verb-like forms in production is evident here, we designed the stimuli using more fine-grained categories to determine whether listeners make use of any finer distinctions. Thus, we included the following grammatical categories: progressive verbs, gerunds, adjectives, nouns, and two types of pronouns: the two-syllable pronouns *something* and *nothing* (pronoun-2s), and the three-syllable pronouns *anything* and *everything* (pronoun-3s). The proportion of stimuli across grammatical categories was guided by Hazen's (2008) large-scale study of conversational speech to mirror distributions in American English; half of the (ING) words were progressive verbs, and half were split among the other five categories. Table 1 displays the grammatical category distribution of the stimuli, along with the distribution of other conditioning factors sometimes shown to influence (ING) realization: preceding and following phonological environments, and the length of the (ING) word.

There were 46 distinct word types among the 96 sentences (e.g., the word *shopping* was in 4 stimuli, twice as a progressive verb, once as a gerund, and once as an adjective). Some words were used multiple times to allow for an initial exploration of word-specific factors over and above more abstract effects like grammatical category, though we do not address this further in this paper. The Corpus of Contemporary American English (Davies, 2008) was used to calculate frequency measures for the (ING) word, the word following the (ING) word, and their bigram frequency.

To make the experiment suitably short for the Mechanical Turk setting, a subset of sentences (40 in total, blocked into four lists of 10 sentences) was presented to each participant. The first list of 10 sentences was presented to all participants. Then, each of the following three lists pulled randomly from the remaining sentences, such that each stimulus was presented to around 30 participants in total (mean N = 31.3, median N = 27, min N = 16, max N (for the first list) = 83). Testing revealed no correspondence between the number of raters and the median rating for a sentence ($r = -0.009$, $p = 0.93$). Within lists, all sentences were presented in random order.

Procedure

Before the task, participants were given the following instructions: "In English, words that end in *–ing* (for example, *talking*) are sometimes pronounced with an *–in* sound (*talkin*) rather than an *–ing* sound (*talking*). In this survey, you will be asked to provide us your opinions about the pronunciation of these words in some written English sentences." Then, before each of the four lists of sentences, participants saw the following: "Imagine that each sentence below is being spoken out loud. Please select how surprising it would be to hear the underlined word in each sentence pronounced with an *–in* rather than *–ing* ending. Answer by rating on a scale of 1 (not at all surprising to hear *–in*) to 6 (extremely surprising to hear *–in*)." Given that *–ing* is the unmarked form, we did not ask whether participants expected to hear *–ing* or *–in*, but rather opted to ask for a self-reported measure of surprise to hear *–in*.

Participants also answered questions about their familiarity with American English, where they have lived, and several other demographic questions. Between experimental lists, attention filters and other short fillers were used to ensure that participants continued to read and follow the instructions.

Results

Likert ratings of stimuli were subjected to statistical analysis via linear mixed-effect modeling[1] (using the lmerTest package for R; Kuznetsova et al., 2014; this package is used for all of the statistical models in this paper). Modeling tested for the maximal random effects structure that improved models by likelihood ratio comparisons and did not lead to convergence errors (using the 'bobyqa' optimizer). Random intercepts were considered for participant, sentence, and (ING) word. All factors described above (under Stimulus design) were tested individually as fixed main effects. Continuous predictors (number of syllables of (ING) word and all frequency measures) were centered around their means. Categorical predictors were dummy coded, with progressive forms as the reference level for grammatical category (since progressives are the most common form in the stimuli and since progressives are expected to show the least *–ing* bias).

Model building proceeded in a manual, step-up fashion, where each factor was tested one at a time using likelihood ratio testing, and added to the model if it significantly improved the model, building up to the most complex model. For random effects, the best model included random intercepts for participant and for word. For fixed effects, all reasonable two-way interactions were examined, though no interactions emerged as significant. All significant fixed effects were included as by-participant random slopes, after assessing that they improved the model. Grammatical category, (centered) number of syllables in word, and (centered) log frequency arose as significant predictors. Table 2 displays an analysis of deviance table for the model, calculated using the Anova function in the car package for R (Fox & Weisberg, 2011). For discussions of statistical results throughout the paper, model estimates and *p* values for categorical factors with more than two levels were calculated through post hoc comparisons of contrasts, with the glht function of the multcomp package in R using the normal distribution to evaluate significance (cf. Hothorn, Bretz, & Westfall, 2008). Fig. 1 displays the raw ratings of surprise to hear *–in* for grammatical category, summarized by-item, demonstrating participants' gradient sensitivity to grammatical category realizations according to production norms.

The model confirmed that pronoun-3 forms (*anything* and *everything*, *M* = 4.76) were rated the most surprising to be realized with *–in*, significantly more so than all other grammatical categories other than nouns (vs. progressive: $\beta = 1.732$, $p < 0.001$; vs. pronoun-2: $\beta = 1.550$, $p = 0.010$; vs. gerund: $\beta = 1.464$, $p = 0.002$; vs. adjective: $\beta = 1.094$, $p = 0.025$). While adjectives received the second highest average raw rating (*M* = 4.01), nouns (*M* = 3.52) were determined by the model to be rated as more surprising, significantly higher than adjectives ($\beta = 0.739$, $p = 0.021$), gerunds ($\beta = 1.109$, $p < 0.001$), pronoun-2s ($\beta = 1.195$, $p = 0.019$), and progressives ($\beta = 1.377$, $p < 0.001$). Adjectives were significantly more surprising than progressives ($\beta = 0.637$, $p < 0.001$). Gerunds (*M* = 2.90), progressives (*M* = 2.67), and pronoun-2s (*M* = 2.38) were rated least surprising and were not significantly different from one another. Model results also indicate that longer words were rated as more surprising ($\beta = 0.761$, $p < 0.001$), and that comprehenders were sensitive to word frequency, rating *–in* to be significantly less surprising as log (ING) word frequency increased ($\beta = -0.164$, $p < 0.001$).

Table 2
Experiment 1 (text) ratings: Analysis of deviance table.

	χ^2	DF	$\Pr(>\chi^2)$
Grammatical category	60.529	5	< 0.00001
Word length (in syllables, centered)	24.867	1	< 0.00001
Word frequency (logged & centered)	21.495	1	< 0.00001

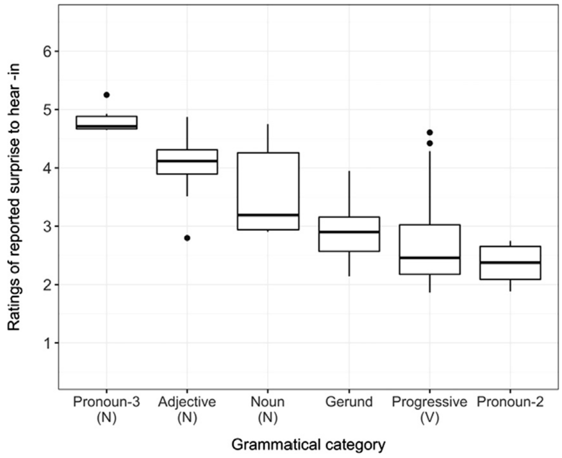

Fig. 1. Experiment 1 (text): Boxplot of participants' self-reported ratings of surprise to hear *–in*, by grammatical category, summarized by item. Rating scale: 1 = Least surprised to hear *–in*, 6 = Most surprised to hear *–in*. Grammatical categories arranged according to raw mean rating. The (N) label marks categories expected to pattern in noun-like ways; (V) marks the verb-like category.

Discussion

Results showed that responses closely reflected grammatical category patterns expected from the descriptive literature on (ING) production patterns, with noun-like forms (pronoun-3s, adjectives, and nouns) showing a large difference from verb-like forms. Gerunds were, as expected, intermediate, and pronoun-2s, aligning with their high rates in production, did not pattern with the other noun-like forms. In addition to reflecting the robust effect of grammatical category, comprehenders' ratings also showed evidence of sensitivity to factors that the production literature has either somewhat neglected (i.e., word length, Kendall, 2013; Tagliamonte, 2004) or has been less definitive about (i.e., lexical frequency, Abramowicz, 2007; Forrest, 2015).

Although this experiment intended to probe comprehenders' intuitions about (ING) production in the absence of any speech input, it may be that in doing the task participants produced the sentences themselves (either aloud or subvocally), and based their responses on their own productions of the stimuli. While we cannot discount this possibility, the task still privileges signal-based information less than if participants had been explicitly presented with an acoustic production

[1] Cumulative logit mixed-effect modeling, which would also be appropriate for Likert-like dependent variables, obtains substantively equivalent results. Linear models are used here to allow for the presentation of results to match the models used throughout the rest of the paper.

to evaluate (as in Experiments 2 and 3).

Given these results, that comprehenders demonstrated sensitivity to (ING)'s grammatical category conditioning in text, we now ask whether listeners make use of the same information when perceiving spoken language.

Experiment 2: Spliced spoken productions

Compared to Experiment 1, Experiment 2 uses more indirect measures to index expectations: listeners' accuracy and reaction times (RTs) when classifying whether an (ING) word in a spoken sentence was produced with *–ing* or *–in*. Since (as discussed above) different realizations of (ING) do not frequently result in minimal pairs that would be confusable in context, more traditional approaches like minimal pair classification tasks are not available when testing (ING) realization. Instead, we ask listeners to classify which realization they heard. In spoken language, will listeners' responses in the classification task show any evidence of the production conditioning knowledge demonstrated in Experiment 1?

We expect that if listeners do weight top-down, grammatical information for use in spoken language comprehension, their accuracy and reaction times in classifying the (ING) variant should index their sensitivity to this information. Since it is not a common paradigm, we describe our expectations about the variant classification task here. We assume that participants will bring their canonical form knowledge to the task and expect to hear *–ing* more than *–in* because it is less marked (probably particularly so in listeners' mental models of speech in a laboratory setting). Consequently, participants' optimal strategy is to listen for occurrences of *–in*, rendering the classification task in effect an *–in* monitoring task. Listeners' *–ing*-bias and subsequent *–in*-monitoring strategy have several implications for our predictions about the effects of realization. First, in terms of reaction time, we expect that listeners will be generally faster for *–in* realizations than *–ing* realizations as a consequence of the optimal task strategy being *–in*-monitoring. Second, in terms of accuracy, we expect that listeners will be more accurate for *–ing* because their overall *–ing*-bias will lead them to over-guess *–ing*. Experiment 1's finding that comprehenders expected noun-like grammatical categories to have a stronger *–ing* bias suggests that listeners' expectations about realizations will be conditioned by grammatical categories. Thus, the crucial measure of interest in Experiments 2 and 3 to demonstrate listeners' sensitivity to conditioning factors is the interaction between grammatical category and realization. Here we expect the two patterns described above to be strengthened for grammatical categories where listeners most strongly expect *–ing* (noun-like categories). Specifically, for reaction time, listeners should notice the violation of expectation more quickly in cases when the *–ing* bias is stronger and accordingly respond even faster in those cases for *–in* than *–ing* realizations. For accuracy, listeners' overall bias toward *–ing*, compounded with encountering a grammatical category with strong *–ing* expectations, means that listeners will over-select *–ing* the most for noun-like categories and thus be even more accurate in those cases when the actual realization is *–ing* and less accurate when it is *–in*.

Method

Participants

Participants in Experiment 2 were 100 undergraduates from the University of Oregon's Psychology and Linguistics Subject Pool who received partial course credit for their time. One additional participant was run but was excluded because they recognized some of the speakers' voices. Participants reported no uncorrected hearing loss. All participants were native English speakers or at least highly familiar with American English, having learned English at age 5 or younger.

Stimulus design

The same 96 stimuli described in Experiment 1 were used in Experiment 2. A larger set of 141 initial sentences had been created for potential use. Four female native English speakers recorded the 141 stimuli in a sound-attenuated booth using a Shure SM93 microphone and a Marantz PMD-661 recorder. Two of the speakers were from Southern California (ages 18 and 23), and two were from Oregon (both aged 18). Speakers were told to produce the sentences as naturally as possible, and all were aware of and able to produce the difference between *–ing* and *–in* forms. Each speaker first read all 141 sentences as they were displayed on the screen in their *–ing* form (*I prefer swimming in the ocean*), and then read all sentences in their *–in* form (*I prefer swimmin' in the ocean*, with an apostrophe after the *–in*). We refer to these two sentence types as "frames" hereafter. This ordering was used to avoid contrastive focus effects that may result from reading the same sentence twice in a row, and to avoid an abundance of errors in (ING) realization that may result from randomizing the frame types.[2]

Twenty-four unique sentences were then selected for each speaker based on naturalness, fluency, and prosodic similarity across frames (4 speakers × 24 sentences = 96 total sentences). One additional sentence was selected from each speaker's productions for use in practice trials. Stimuli were RMS amplitude normalized to 70 dB. Then, a splicing procedure (following Campbell-Kibler, 2007; Labov et al., 2011) was conducted to counterbalance the "frame" and the "realization" of each stimulus across participants, creating four versions of each sentence (4 versions × 96 sentences = 384 total stimuli); see Fig. 2 for an illustration of stimulus design in Experiments 2 and 3. Splicing was done even for the frame-realization matched sentences so that all stimuli in Experiment 2 contained equivalent manipulation.[3]

The (ING) words in the stimuli were then analyzed acoustically, both to ensure that speakers were not differentially cuing the (ING) forms for our factors of interest (e.g., by hyperarticulating *–in* forms in grammatical categories most favoring *–ing*) and to provide acoustic and temporal measurements for testing as independent variables in our later analysis. A summary of these acoustic measures and their analysis is provided in Appendix B. Analysis of the acoustic measures confirmed expectations about phonetic differences between *–in* and *–ing*: *–ing* forms were longer, with longer, more tensed (higher and fronter) vowels. Importantly, however, no interactions between grammatical category and realization occurred in the stimuli, confirming that the acoustics of the (ING) in the stimuli were not differentially predicted by grammatical category.

Each participant heard all 96 sentences, in random order. For each sentence, half of the participants heard each frame and, for each frame, half of the participants heard each realization, resulting in 25 participants rating each of the four frame-realization versions of each sentence. This counterbalancing was done within talker, such that listeners heard 12 *–ing* and 12 *–in* realizations (for each, six were originally an

[2] Although using spontaneous productions as stimuli would be ideal, read speech was used because of our interest in comparing across sentence frames, and with the results of Experiment 1. However, there is precedent for using read speech as stimuli in studies of listener sensitivity to (ING) (e.g., Labov et al., 2011), and Tamminga (2017) found no significant differences in listeners' social perceptions of (ING) for read versus spontaneous stimuli.

[3] To conduct the splicing, the (ING) realizations in each of the sentential frames were segmented by hand based on auditory and spectrographic cues in Praat (Boersma and Weenink, 2015), and the boundaries of the realization were selected at the nearest zero-crossings in the acoustic waveform. A new realization of the (ING) form, selected from a copy of the sentence recording using the same segmentation process, was pasted in place of the original realization, once for *–in* and once for *–ing*, for each frame. For the set of stimuli that contained matching (ING) realizations to the original frames (i.e., *–ing* realization pasted into the *–ing* frame and *–in* pasted into the *–in* frame), care was put into ensuring that identical zero-crossings were not selected as this would result in the recreation of the exact original frame, and slight (sometimes as little as a few milliseconds) differences were desired to ensure that any artifacts created in the splicing of mismatching realizations and frames would also occur in the matched cases.

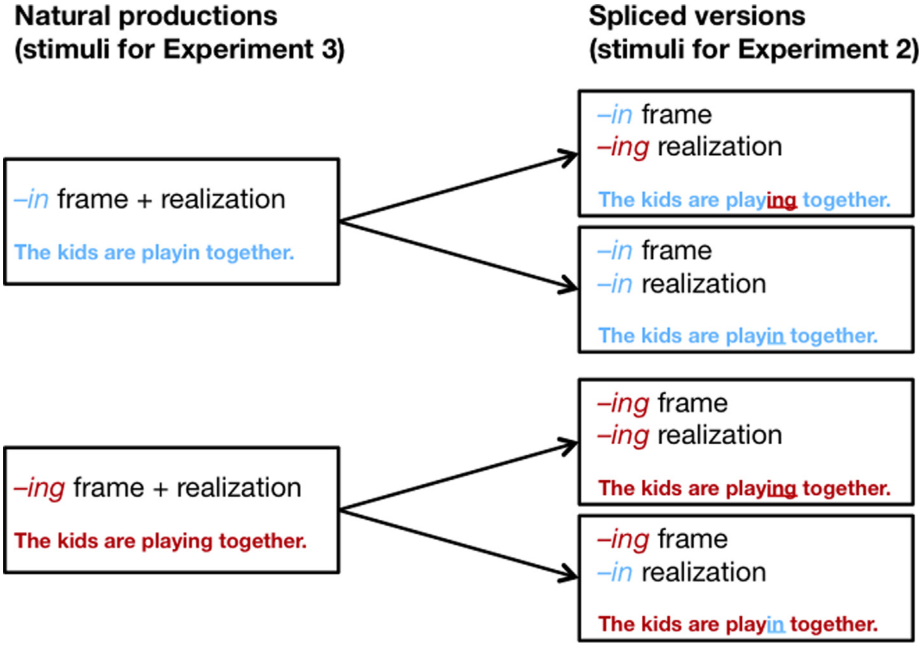
Natural productions (stimuli for Experiment 3)
Spliced versions (stimuli for Experiment 2)
–in frame + realization
The kids are playin together.
–in frame
–ing realization
The kids are playing together.
–in frame
–in realization
The kids are playin together.
–ing frame + realization
The kids are playing together.
–ing frame
–ing realization
The kids are playing together.
–ing frame
–in realization
The kids are playin together.

Fig. 2. Schematic of stimulus design for Experiments 2 and 3.

–ing frame and six were originally an *–in* frame) from each talker. Thus, for every listener, each talker's overall rate of (ING) realization was 50% *–ing*, 50% *–in*. (Listeners were not told that stimuli would be balanced between *–in* and *–ing* realizations and given an *–ing* bias, as described above, listeners likely expected a higher proportion of *–ing*.)

Procedure

Participants completed the (ING) classification task individually seated in a sound-attenuated booth in front of a PST Serial Response Box, wearing Sennheiser HD-202 headphones. The task was presented using E-Prime 2.0 (Psychology Software Tools). Participants were first reminded that English speakers sometimes produce words that end in (ING) as *–ing* (as in *talking*), and sometimes as *–in* (as in *talkin*). Then, participants were informed that they would be listening to a series of sentences that each contained an (ING) word. They were asked to: "Press the *–ing* button when you hear a word ending in *–ing* and the *–in* button when you hear a word ending in *–in*", before being presented with four practice trials. Participants were encouraged to respond as quickly as possible without sacrificing accuracy, and were explicitly instructed that they could respond before the sentence was over. The correspondence between the position of the buttons and the possible responses, *–ing* and *–in*, was counterbalanced across participants. Following the classification task, all participants completed a standard demographic and language background questionnaire.

Results

9600 responses were collected in total (96 items × 100 participants). Before analysis, data points where participants responded before the *–ing* or *–in* onset were removed (N = 61). RTs were then transformed to natural log values and responses greater than ± 2.5 standard deviations from each participant's mean log RT were trimmed from the dataset (N = 122). Combined, these procedures resulted in the removal of 1.9% (N = 183) of the data. Table 3 summarizes participants' accuracy and log reaction times by (ING) realization and grammatical category.

Accuracy

Overall, participants responded with the accurate realization 80.4% of the time (N = 7572). To evaluate participants' patterns of errors, statistical analysis of item accuracy as the dependent variable was conducted via mixed-effect logistic regression. The modeling procedure followed that described for Experiment 1, testing all relevant fixed effects and the maximal random effects structure the model would support without convergence errors and that improved the model. Fixed effects tested included all those outlined in Experiment 1, the acoustic predictors in Appendix B, plus location in the sentence of the (ING) form (measured as a proportion of the duration of each sentence), and trial number (the serial position of the stimulus in the experiment for a participant). Two-way interactions were tested for all reasonable pairings of factors and three-way interactions were tested for sets of factors deemed especially important; no three-way interactions were significant. The final model included random intercepts for participant and sentence, but no random slopes.[4] Table 4 displays an analysis of deviance table for the fixed effects that emerged in the best model. As was done for Experiment 1, model estimates and *p* values for categorical factors with more than two levels were calculated through post hoc comparisons of contrasts.

The model results (Table 4) indicate that a number of factors significantly influenced participants' accuracy. Participants were significantly less accurate for *–in* than *–ing* realizations ($\beta = -1.264$, $p < 0.001$) in line with our predictions. The significant interaction between realization and frame indicated improved accuracy for *–in* realizations when the frame was *–in* ($\beta = 1.029$, $p < 0.001$). Main effects for RT and the duration of the (ING) form indicate that participants were significantly more accurate when they were faster ($\beta = -0.756$, $p < 0.001$) and when the (ING) form was longer ($\beta = 1.111$, $p < 0.001$). Individual post hoc comparisons for the main effect of following environment were not significant, but the interaction between the (ING) realization and the following environment indicates that participants were more accurate when *–in* realizations were followed by an alveolar form, as would be expected (vs. velar: $\beta = 0.721$, $p < 0.001$; vs. other: $\beta = 0.533$, $p < 0.001$). For grammatical category, no individual post hoc comparisons reached significance as main effects, but most crucially, grammatical category significantly

[4] Models including random slopes that obtained convergence errors nonetheless yield comparable results to the models reported here for the factors of interest. This is the case for all such models reported later in the paper as well.

Table 3
Experiment 2 (spliced audio): Accurate responses/total responses (and percent correct) and (natural log) reaction times, by grammatical category and realization.

			Grammatical category						*Overall*
			Pronoun-3	Adjective	Noun	Gerund	Progressive	Pronoun-2	
Realization	*–ing*	N acc/total (%)	250/292 (85.6)	660/765 (86.3)	376/435 (86.4)	469/547 (85.7)	2011/2380 (84.5)	252/290 (86.9)	*85.3%*
		Mean log RT (sd)	6.98 (0.62)	6.96 (0.59)	6.97 (0.65)	6.90 (0.51)	6.87 (0.57)	6.83 (0.62)	*6.90*
	–in	N acc/total (%)	166/292 (56.1)	576/765 (75.3)	357/429 (83.2)	415/544 (76.3)	1813/2380 (76.2)	227/294 (77.2%)	*75.5%*
		Mean log RT (sd)	6.61 (0.58)	6.90 (0.60)	6.82 (0.65)	6.80 (0.52)	6.78 (0.53)	6.73 (0.58)	*6.79*
Overall		*% accurate*	*70.7%*	*80.8%*	*84.8%*	*81.0%*	*80.3%*	*82.0%*	*80.4%*
		Mean log RT	*6.83*	*6.93*	*6.90*	*6.85*	*6.83*	*6.78*	*6.90*

Table 4
Experiment 2 (spliced audio) accuracy: Analysis of deviance table.

	χ^2	DF	$\Pr(>\chi^2)$
Realization	96.632	1	< 0.00001
Frame	0.018	1	0.89200
Grammatical category	16.202	5	0.00629
Following environment	13.789	2	0.00101
RT (logged & centered)	177.673	1	< 0.00001
(ING) Duration (logged & centered)	44.389	1	< 0.00001
Realization × Frame	80.956	1	< 0.00001
Realization × Gram. Cat.	36.690	5	< 0.00001
Realization × Fol. Env.	24.281	2	0.00001

Table 5
Experiment 2 (spliced audio) reaction time: Analysis of deviance table.

	χ^2	DF	$\Pr(>\chi^2)$
Realization	42.002	1	< 0.00001
Frame	5.322	1	0.02106
Grammatical category	19.115	5	0.00183
Location of (ING)	94.566	1	< 0.00001
Trial number	16.582	1	0.00005
(ING) Duration (logged & centered)	18.729	1	0.00002
Vowel F2 (Bark & centered)	17.178	1	0.00003
Realization × Frame	17.428	1	0.00003
Realization × Gram. Cat.	30.465	5	0.00001

interacted with realization in influencing accuracy patterns. Participants were significantly less accurate for pronoun-3 forms than all other grammatical categories, when they were realized as *–in* (vs. pronoun-2: $\beta = -1.207$, $p < 0.001$; vs. progressive: $\beta = -1.466$, $p < 0.001$; vs. gerund: $\beta = -1.313$, $p < 0.001$; vs. noun: $\beta = -1.568$, $p < 0.001$; vs. adjective: $\beta = -1.346$, $p < 0.001$). No other grammatical category comparisons reached significance.

Given that participants have a strong expectation for *–ing* overall, it is perhaps not surprising that the only grammatical category surfacing as significant in interaction with realization, pronoun-3, is the one that showed the biggest difference from progressives in Experiment 1. This overall bias to respond *–ing* makes accuracy not particularly informative, and thus we expect that reaction times will be a more sensitive measure to differences between grammatical categories.

Reaction times

RTs were measured relative to the onset of the vowel in the (ING) portion of the critical word and transformed to natural log values. Log RTs were then centered around the mean for modeling and inaccurate responses were removed for RT analysis, leaving N = 7572 responses. The mean log RT for accurate responses was 6.852 (St. Dev. = 0.576). The data were subjected to mixed-effect linear regression modeling with centered log RT as the dependent variable. The modeling procedure followed that described for the accuracy analysis, testing all factors described above and in Appendix B, although we only discuss significant factors. The final model included random intercepts for participant and sentence. An analysis of deviance table for the best model is shown in Table 5 and, again, model estimates and *p* values for grammatical category components of the model were calculated through post hoc comparisons of contrasts.

As indicated in Table 5, a number of significant effects influence participants' RTs. For the main effect of grammatical category, adjectives and nouns elicited significantly longer log RTs than progressives (adjective: $\beta = 0.147$, $p = 0.033$; noun: $\beta = 0.239$, $p = 0.002$) and nouns elicited marginally longer log RTs than gerunds ($\beta = 0.214$, $p = 0.059$); no other grammatical category main effect differences approached significance. For location in sentence, participants responded more quickly to (ING) words occurring later in the sentence ($\beta = -0.001$, $p < 0.001$). For trial number, participants took longer as they proceeded through the experiment ($\beta = 0.001$, $p < 0.001$). Also, longer (ING)s led to significantly faster log RTs ($\beta = -0.149$, $p < 0.001$), and more fronted (ING) vowels (higher F2) led to slower log RTs ($\beta = 0.027$, $p < 0.001$). Realization and frame contribute significantly to the model as main effects (as evidenced in Table 5), but their significant influences on log RTs only arise in interactions (main effects: realization of *–in*: $\beta = -0.015$, $p = 0.46$; frame of *–in*: $\beta = 0.017$, $p = 0.24$).

Crucially, the realization of the (ING) form significantly interacted with the (ING) word's grammatical category, visualized in Fig. 3. For *–in* realizations, as predicted, nouns and pronoun-3s (both noun-like categories) were significantly faster than progressives (noun: $\beta = -0.125$, $p = 0.009$; pronoun-3: $\beta = -0.173$, $p = 0.013$) and were marginally faster than gerunds (noun: β = -0.118, $p = 0.099$; pronoun-3: $\beta = -0.165$, $p = 0.055$), a finding in line with gerunds' status as in-between nouns and verbs in their patterns for (ING). Unexpectedly, nouns ($\beta = -0.174$, $p < 0.001$) and pronoun-3s ($\beta = -0.222$, $p = 0.001$) were significantly faster for *–in* than adjectives as well, a

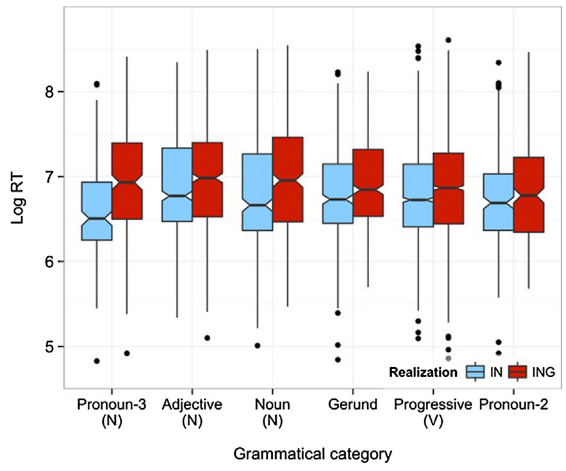

Fig. 3. Experiment 2 (spliced audio): Boxplot of listeners' log reaction times, demonstrating interaction between realization heard (*–in* or *–ing*) and grammatical category. Grammatical categories arranged according to rank ordering obtained in Experiment 1. The (N) label marks categories expected to pattern in noun-like ways; (V) marks the verb-like category.

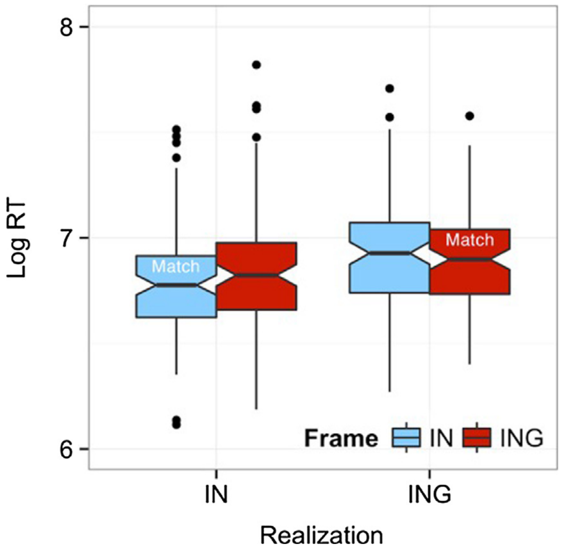

Fig. 4. Experiment 2 (spliced audio): Boxplot of listeners' log reaction times, summarized by item, demonstrating interaction between realization (*–in* or *–ing*) and frame (*–in* or *–ing*).

surprising outcome given the fact that adjectives patterned with other noun-like forms in Experiment 1. Finally, pronoun-2s for *–in* were significantly faster than adjectives (β = −0.141, *p* = 0.0498) but not significantly different than any other grammatical category. No other contrasts approached significance.

Finally, in line with accuracy patterns, the interaction between (ING) realization and original (ING) frame was significant: participants were significantly faster for *–in* realizations when the original sentence was also produced in an *–in* frame (β = −0.088, *p* < 0.001) – i.e., where the frame and realization matched. This interaction is depicted in Fig. 4.

Discussion

The results of Experiment 2 indicate that listeners did demonstrate sensitivity to grammatical category expectations when asked to categorize realizations of (ING). Significant interactions emerged between grammatical category and realization, where listeners responded to *–in* realizations more quickly for grammatical categories whose *–ing* bias was stronger (noun and pronoun-3 forms, though not adjectives). Listeners were also least accurate when the category least expected as *–in* in Experiment 1, pronoun-3, was realized with *–in*. That is, participants were fastest to respond, and least accurate, when the grammatical category would have led them to most strongly expect *–ing*, but instead they were presented with *–in*. These patterns demonstrate that listeners do weight their expectations based on production patterns in auditory contexts.

Further, for RTs, the significant effect of (ING) location in the stimulus sentence indicates that when listeners had heard more of the sentence before they heard the (ING) word, they were faster to respond: when the (ING) word was later in the sentence they could make use of more information in the signal (including everything from acoustic to syntactic and semantic information) to predict when an (ING) word was coming up. Also, the significant trial number effect showed that participants had slower RTs as the experiment went on, a finding we return to below. Finally, significant effects for aspects of the (ING) productions—the duration of (ING), the F2 of the (ING) vowel (for RTs), and the phonological environment following the (ING) (for accuracy)—indicate that, unsurprisingly, listeners are influenced by direct acoustic characteristics of the stimuli. Importantly, these acoustic factors did not interact with grammatical category and appeared in addition to, not in place of, the influence of other factors like grammatical category and realization.

We now explore in further detail two potential factors that may modulate the effects demonstrated here: how classifiable each (ING) word is in isolation as *–ing* or *–in*, and the role of matching or mismatching frames and realizations. We consider the first factor by briefly presenting a follow-up experiment, before describing the second, which motivates Experiment 3.

Role of isolated word-level differences in perception of realization

One possible limitation of these results is the potential for different stimulus items to contain differing strengths of acoustic cues favoring the classification of a token as *–ing* versus *–in*. Our acoustic analyses of the stimuli (in Appendix B) support that acoustic properties of (ING) were not behind the critical interaction between grammatical category and realization. However, the realization factor included in the model was based on speakers' *intended* realizations, and it could be the case that, even in isolation, listeners would not always classify an (ING) according to the speaker's intended realization. Since intended realization was a factor in the modeling for Experiment 2, and accuracy of responses was calculated in reference to intended realizations, we wanted to determine the extent that speakers' intended realizations, in fact, correlated with listeners' percepts.

To do this, we asked a separate group of listeners (N = 50, from the same population) to classify realizations of these same (ING) word tokens when the (ING) word was isolated from its sentential context. We excised each (ING) word from each of our 96 sentences, and used those individual words as stimuli for this new group of participants. In this task, participants were also asked to classify each (ING) word as either *–in* or *–ing*. Half of the participants heard the speaker's intended *–ing* form of a word, and the other half heard the intended *–in* form. We used half of the stimuli (192 items) from Experiment 2, choosing the word from the original *–ing* frame for half the words, and the original *–in* frame for the other half, again with two realization versions for each word. Results showed that participants responded in accordance with the speaker's intended realization of each word 84.6% of the time (range across all words = 63.3–98.0% accurate).

To demonstrate that our effect of interest in Experiment 2 cannot be accounted for by systematic differences in the ease of classification of particular tokens in isolation, we re-analyzed the reaction time data from Experiment 2, from the half of the participants in Experiment 2 who heard the 192 items included in the follow-up experiment. (We focus on the RT results here for sake of space and since that model was more informative.) To do this, we reexamined the RT model from Experiment 2 and replaced the realization term (*–ing* or *–in*, based on intended production) with the proportion heard as *–in* term (a new, continuous measure calculated from participants' responses in the follow-up experiment). The new model fit the data quite well, with all of the factors from the main analysis remaining in the best model. As with realization in the main model (Table 5 above), proportion heard as *–in* significantly interacted with sentence frame (*–in*: β = −0.109, χ^2=6.85, *p* = 0.009) and with grammatical category (χ^2=16.80, *p* = 0.005), as a predictor of log RT, with the same grammatical categories responsible for the effect as reported in Experiment 2. Thus, our main finding of interest, that listeners are faster to identify *–in* realizations of grammatical categories for which they have strong *–ing* expectations, still held even when taking into account item-level variability in the inherent identifiability of realizations as *–ing* or *–in*.

Role of (in)consistent bottom-up cues

The splicing methodology was adopted in Experiment 2 to isolate the effect of the realization of (ING) itself on classification decisions (following, e.g., Campbell-Kibler, 2007; Labov et al., 2011). This approach ensured that acoustic cues beyond the critical element of the stimulus were held constant, but also meant that half of the stimuli contained an (ING) realization embedded in a sentential frame that was

Table 6
Experiment 3 (natural audio): Accurate responses/total responses (and percent correct) and (natural log) reaction times, by grammatical category and realization.

			Grammatical category						*Overall*
			Pronoun-3	Adjective	Noun	Gerund	Progressive	Pronoun-2	
Realization	*–ing*	N acc/total (%)	124/143 (86.7)	342/383 (89.3)	189/215 (87.9)	238/271 (87.8)	1044/1185 (88.1)	128/143 (89.5)	*88.2%*
		Mean log RT (sd)	6.89 (0.61)	6.82 (0.63)	6.83 (0.69)	6.87 (0.52)	6.81 (0.56)	6.93 (0.57)	*6.83*
	–in	N acc/total (%)	103/148 (69.6)	306/388 (78.9)	187/216 (86.6)	226/272 (83.1)	984/1194 (82.4)	112/147 (76.2)	*81.1%*
		Mean log RT (sd)	6.56 (0.53)	6.75 (0.60)	6.73 (0.66)	6.79 (0.57)	6.68 (0.53)	6.70 (0.61)	*6.70*
Overall		*% accurate*	*78.0%*	*84.0%*	*87.2%*	*85.5%*	*85.2%*	*82.8%*	*84.7%*
		Mean log RT	*6.74*	*6.79*	*6.78*	*6.83*	*6.75*	*6.82*	*6.77*

originally produced with the opposite realization, and thus presumably contained acoustic cues consistent with that original realization. (While we have considered acoustic characteristics of the stimuli's (ING) forms in our analyses, we have not included potential cues from elsewhere in the stimuli.) As a result, listeners were regularly confronted with acoustic cues in the sentence's frame that would not reliably predict the expected realization of the (ING) word, and several results suggest that participants were indeed sensitive to this. For example, the significant interaction of frame and realization indicated that listeners' responses were more accurate when frame and realization matched and were fastest overall when the frame was *–in* and the realization was *–in*. The frame provided listeners with advance evidence that an *–in* was upcoming, and that evidence was confirmed in the realization, so participants responded quickly (Fig. 4). Thus, especially for *–in* frames, there was evidence of a frame/realization mismatch effect. Further, the significant effect of trial number indicates that listeners took longer to respond as the experiment progressed; we interpret this result as suggesting that listeners realized that their expectations based on the sentence frames were not always reliable and reweighted their use of expectation-based prediction over the course of the experiment.

These patterns are an important, novel demonstration of listeners' sensitivity to phonetic congruence on a sentential level, and are reminiscent of prior findings on the word level (Sumner et al., 2013; Sumner, 2013). Sumner and colleagues previously observed that congruence between a particular pronunciation variant (e.g., realizations of word-medial /t/) and the surrounding word-level phonetic cues (e.g., produced either "carefully" or "casually") influences the interpretation of that pronunciation variant. That listeners in our Experiment 2 demonstrated sensitivity to the original frame in which the sentence was produced, and that neither acoustic measures of the (ING) nor the follow-up experiment on classification of individual (ING) words in isolation (both aspects of the (ING) words themselves) fully accounted for RTs, suggests that cues in the signal that were not immediately proximate to the variable contributed to expectations about the variable's realization. Thus, Experiment 3 used stimuli with congruent frames and realizations.

Experiment 3: Natural spoken productions

In real life speech settings, speakers reliably produce consistent cues, unlike the situation in Experiment 2. Therefore, Experiment 3 uses unmanipulated, natural productions as stimuli in a variant classification task. Taken together with results from Experiments 1 and 2, the results of this experiment address RQ2, probing when listeners are more or less likely to use expectations about probabilistic conditioning. We predicted that the use of grammatical category information (demonstrated by the interaction between grammatical category and realization) in Experiment 3 should be mitigated in comparison to Experiment 2 because bottom-up cues are more reliable.

Method

Participants

Participants in this experiment were 50 undergraduates from the University of Oregon's Psychology and Linguistics Subject Pool who received partial course credit for their time. No participants had completed any other previous experiments in this study. Five additional participants were run but were excluded due to hearing impairments (N = 4) or being a non-native English speaker (N = 1). Remaining participants reported no uncorrected hearing loss and were native English speakers or at least highly familiar with American English, having learned English at age 5 or younger.

Stimulus design

The stimuli were the same 96 sentences used in previous experiments, with two versions of each sentence, *–ing* and *–in* (realization and frame are equivalent here; refer to Fig. 2). Sentences were presented to participants as they were produced by speakers, without splicing or other manipulation aside from RMS amplitude normalization to 70 dB. As with Experiment 2, these stimuli were also analyzed acoustically (summarized in Appendix B). Results from acoustic analyses were in line with those from Experiment 2: *–ing* forms were longer, with longer, more tense vowels, but there were no significant interactions between grammatical category and realization, confirming that the (ING) realizations were not differentially cued by grammatical category.

Procedure

The procedure was identical to that used in Experiment 2; the only difference between Experiments 2 and 3 was the stimuli.

Results

4800 responses were collected in total (96 items × 50 participants). Responses before the (ING) onset were removed (N = 26). Responses ± 2.5 standard deviations from each participant's mean log RT were trimmed (N = 69), resulting in the removal of 2.0% (N = 95) of the data. Table 6 summarizes participants' accuracy and log reaction times by (ING) realization and grammatical category.

Accuracy

Overall, participants responded with the accurate realization 84.7% (N = 3,983) of the time. Accuracy was significantly higher than in Experiment 2 ($\chi^2 = 37.8$, $p < 0.001$).

Statistical analysis followed the same approach as described for Experiment 2. The final model included random intercepts for participant and item. An analysis of deviance table for the best model is presented in Table 7.

The model confirms that *–ing* realizations were significantly more accurate than *–in* realizations overall (β = −1.231, $p < 0.001$), as in Experiment 2. Main effects for RT and the duration of the (ING) form indicate that participants were significantly more accurate when they were faster (β = −0.647, $p < 0.001$) and when the (ING) form was longer (β = 0.736, $p = 0.002$). Unlike in Experiment 2, the intensity of

Table 7
Experiment 3 (natural audio) accuracy: Analysis of deviance table.

	χ^2	DF	Pr(> χ^2)
Realization	38.263	1	< 0.00001
Grammatical category	6.972	5	0.22271
Following environment	7.888	2	0.01937
RT (logged & centered)	56.220	1	< 0.00001
(ING) Duration (logged & centered)	9.217	1	0.00240
(ING) Intensity (centered)	5.914	1	0.01502
Realization × Gram. Cat.	11.964	5	0.03528
Realization × Fol. Env.	29.258	2	< 0.00001
Realization × (ING) Intens.	7.420	1	0.00645

the (ING) also significantly influenced accuracy, with participants less accurate for *–in* forms with higher intensity ($\beta = -0.070$, $p = 0.006$). Participants were once again influenced by the interaction between the (ING) form and its following environment (alveolar vs. velar: $\beta = 1.307$, $p < 0.001$; other vs. velar: $\beta = 0.995$, $p < 0.001$). Grammatical category was not significant as a main effect but again was significant in interaction with realization. However, the post hoc comparison between pronoun-3s (the grammatical category rated with highest surprise to hear *–in* in Experiment 1) and progressives for *–in* forms was the only contrast to obtain significance ($\beta = -1.150$, $p = 0.026$).

Reaction times

The RT analysis for Experiment 3 followed the logic and procedure of Experiment 2, with only the accurate responses (N = 3983) examined. Overall, the mean log RT for accurate responses in Experiment 3 was 6.771 (St. Dev = 0.579). This was significantly faster than the reaction times in Experiment 2 ($M = 6.852$; $t = 7.137$, $p < 0.001$). Centered log RTs were submitted to mixed-effect linear regression. The best model included random intercepts for participant and sentence. An analysis of deviance table for the best model is presented in Table 8.

Similar to Experiment 2, participants were faster when the (ING) form appeared later in the sentence ($\beta = -0.001$, $p < 0.001$) and when the (ING) was longer ($\beta = -0.145$, $p < 0.001$), and they were slower when the vowel in (ING) was more front ($\beta = 0.038$, $p < 0.001$). Although grammatical category significantly contributed to the model as a main effect, post hoc tests indicate that no contrasts obtain significance (due to conservative criteria for multiple comparisons). Significant interactions occurred between (ING) realization and grammatical category, between (ING) realization and the phonological environment following the (ING), and between (ING) realization and the F2 of the vowel in the (ING). The interaction of greatest interest, between realization and grammatical category, was driven by slower RTs for adjectives in *–in* compared to several other categories, a pattern unexpected from prior work and Experiment 1's results (adjective vs. progressive: $\beta = 0.127$, $p = 0.024$; adjective vs. pronoun-2: $\beta = 0.198$, $p = 0.046$; adjective vs. pronoun-3: $\beta = 0.248$, $p = 0.010$). This mitigated interaction between realization and grammatical category is evident in Fig. 5, where listeners' responses to *–in* realizations were not faster for noun-like categories compared to verb-like categories, as they were in Experiment 2 (Fig. 3).

Table 8
Experiment 3 (natural audio) reaction time: Analysis of deviance table.

	χ^2	DF	Pr(> χ^2)
Realization	23.020	1	< 0.00001
Grammatical category	12.038	5	0.03427
Location of (ING)	41.928	1	< 0.00001
(ING) Duration (logged & centered)	11.967	1	0.00054
Vowel F2 (Bark & centered)	29.680	1	< 0.00001
Following environment	4.496	2	0.10563
Realization × Gram. Cat.	17.170	5	0.00419
Realization × Vowel F2	6.225	1	0.01259
Realization × Fol. Env.	20.294	2	0.00004

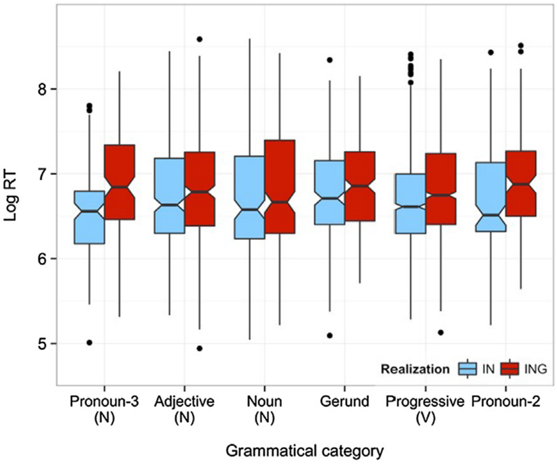

Fig. 5. Experiment 3 (natural audio): Boxplot of listeners' log reaction times, demonstrating mitigated interaction between realization (*–in* or *–ing*) and grammatical category. Grammatical categories arranged according to rank ordering found in Experiment 1. The (N) label marks categories expected to pattern in noun-like ways; (V) marks the verb-like category.

Further, unlike Experiment 2, which obtained no significant RT effects for the phonological environment surrounding the (ING), here participants had faster RTs when the realization was *–in* and the following sound was also an alveolar (vs. velar: $\beta = -0.159$, $p < 0.001$; vs. other: $\beta = -0.129$, $p < 0.001$). Also, the interaction between realization and the F2 of the (ING) vowel indicates that listeners were less fast for *–in* forms with fronter vowels ($\beta = 0.041$, $p = 0.013$). Altogether, these patterns suggest that bottom-up cues may have played a larger role in Experiment 3 than Experiment 2.

Discussion

Participants were faster overall, and more accurate, in Experiment 3 than in Experiment 2, which we suggest is due to the unreliability of evidence from conflicting frames in Experiment 2 as compared to the congruent acoustic cues in Experiment 3. Potentially supportive of this explanation is the observation that the trial number effect in Experiment 2 (where participants had slower RTs in later trials, interpreted as evidence that listeners became aware of incongruence and slowed down to account for it) no longer emerged as predictive in Experiment 3.

Further, grammatical category influenced reaction times for naturally produced sentences less so than it did for the spliced sentences of Experiment 2; listeners in Experiment 3 down-weighted their expectations when the bottom-up information was consistently reliable. Further, significant effects for phonological environment in Experiment 3 suggest that listeners in Experiment 3 were attending more to the bottom-up signal than they were in Experiment 2, consistent with this account. However, evidence of top-down sensitivity is not completely absent since, for instance, grammatical category in interaction with realization had similar effects on accuracy in Experiments 2 and 3, suggesting the potential that listeners may still use these expectations in real-life listening situations.

General discussion

This study tested whether listeners have expectations about the grammatical category constraints that condition the realization of (ING), even though such knowledge may not be very informative for word recognition. In Experiment 1, sentences containing (ING) words were presented to listeners in written form, who rated how surprised they would be to hear each (ING) word pronounced as *–in*. Participants reported the highest amount of surprise to hear *–in* for (ING) words in grammatical categories least favoring of *–in* in production studies (noun-like categories). In Experiment 2, participants classified the realization of (ING) words in spoken versions of the stimuli with (ING) realizations spliced in, and were faster and less accurate in classifying stimuli with realizations that mismatched their expectations in terms of grammatical category. In Experiment 3, with naturally produced versions of the same spoken stimuli, the grammatical category by realization interaction was mitigated and phonological effects arose for RT, suggesting that listeners did not rely on grammatical category expectations as much with reliable bottom-up information.

Thus, returning to RQ1, we found that listeners are indeed sensitive to probabilistic conditioning constraints of (ING) realization, despite the fact that its variation does not readily introduce lexical ambiguity. Of course, given the nature of the tasks used in these experiments, we cannot determine whether listeners actually generate online predictions about how (ING) is likely to be realized during the course of everyday speech perception. In fact, our results with respect to RQ2 show that listeners are less likely to use these expectations with more reliable cues. However, this study provides evidence that the information is nonetheless available for listeners' use.

With regard to RQ2, results across experiments are in line with previous work suggesting that the comprehension system is opportunistic: comprehenders rely on bottom-up acoustic information if it is available and reliable, and on top-down (e.g., grammatical) information if it is not. Experiment 1, using text alone, demonstrated sensitivity to top-down knowledge about the conditioning of (ING) as indexed by main effects of grammatical category and word frequency. However, Experiment 3, using naturalistic audio, found that such top-down information was less important, and bottom-up factors were more predictive. The results of Experiment 2, using spliced audio with inconsistent bottom-up cues, were more mixed, showing evidence that interactions occur between top-down and bottom-up factors (e.g., grammatical category by realization interaction), as well as among bottom-up factors (e.g., frame by realization interaction).

As discussed above, the results of Experiment 2 are in line with recent work by Sumner (2013; Sumner et al., 2013), which demonstrates the important role of congruence between a word's phonetic frame and a pronunciation variant of a segment. We provide evidence for such congruence effects on the sentence level. Participants in Experiments 2 and 3 demonstrated sensitivity to the stylistic coherence of each stimulus, with faster reaction times and higher accuracy in cases when frames and realizations matched, especially in *–in* forms. These results are suggestive that stylistic cues may interact with listeners' expectations about conditioning by grammatical category, and more generally that listeners may rely on many features in the speech signal when making linguistic judgments, possibilities that we are directly investigating (Vaughn & Kendall, 2018).

Theoretical implications

The results of this study have significant implications for theories involving language prediction and expectation in speech perception. We found that, even though monitoring (ING) realization is not strictly necessary for word disambiguation, listeners demonstrate knowledge of its probabilistic conditioning constraints in production. Since monitoring the probabilistic conditioning of every linguistic variable and generating expectations conditioned on infinitely many factors would be computationally intractable, the onus rests on theories to account for the conditions under which the system is sensitive to such variation (cf. MacDonald & Seidenberg, 2006). Many current theories primarily use a feature's utility for linguistic categorization as the criterion to account for when features should be tracked, stored, and/or predicted, in an attempt to optimize recognition or categorization (e.g., Baayen et al., 2013; Kleinschmidt & Jaeger, 2015; Levy, 2008; McMurray & Jongman, 2011). The precise statistics tracked by models will vary depending on what the model is trying to explain (MacDonald & Seidenberg, 2006). Thus, it is not surprising that prior work has not endeavored to include such information: the preponderance of psycholinguistic studies in this area are designed to explain recognition and categorization. Nonetheless, our participants demonstrated grammatical category knowledge about (ING), which is not especially informative about word recognition.

We outline two interpretations of our findings, which are not mutually exclusive. First, these results may lend support to theories using what Huettig (2015) discusses as "dumb" prediction, where statistical regularities are learned between any items that co-occur frequently through simple associative mechanisms (predictive coding models and naïve discriminative learning models, for example, could fall under this broad type of theory). Our results would naturally fall out from models based on this type of prediction, even if the models were not explicitly designed to predict such effects.

The second interpretation of these results is that there *is* some utility to tracking the probabilistic conditioning of (ING) realization, just not utility for word recognition, which means that models should incorporate other kinds of utility. The architecture of many current models (e.g., Baayen et al., 2013; Clark, 2013; Levy, 2008; Kleinschmidt & Jaeger, 2015; McMurray & Jongman, 2011; Sumner, Kim, King, & McGowan, 2014) could take probabilistic constraints that are not informative for word recognition into account. For example, Sumner et al.'s (2014) socially weighted encoding model readily accounts for the frame/realization mismatch effect and the canonical form bias we observed, but not our grammatical category/realization interaction results. The socially weighted encoding model could be augmented to account for our results: rather than making words be the minimal unit of representation, a word *in its usage context* (for example, its grammatical category) may be the minimal unit (as suggested by exemplar theories, e.g., Pierrehumbert, 2002). This modified theory would allow for a particular variant to be the typical realization within a phonetic frame *for a particular usage context (e.g., for a particular grammatical category)*, while preserving the overall mechanisms of the model (see also Bürki et al., 2010; Ranbom & Connine, 2007). Building in expectations about variables' realization patterns by usage context (e.g., grammatical category) may also allow the model to account for other usage-based patterns such as compound words.

Further, it may be that the social utility of monitoring the probabilistic conditioning of a feature contributes to when conditioning information is useful. In many real-life speech situations, comprehension and social evaluation may both demand a listener's attention. For example, interactions where the individual's goal is acceptance by a new social group would likely involve tracking patterns in sociolinguistic variation in addition to recognizing the words being said. For example, using *–in* in an unexpected grammatical category like pronoun-3 (e.g., *everythin*) might send a stronger social message than *–in* in, say, a progressive verb (e.g., *walkin*). In support of this hypothesis, our follow-up work (Vaughn & Kendall, 2018) has found that listeners' ratings of these stimuli as more or less accented is modulated by an interaction between grammatical category and realization, which indicates that conditioning information is informative for social evaluation (see also Bender, 2005). These results, along with recent empirical work demonstrating sensitivity to covariation between social and linguistic patterning of variables (e.g., Docherty, Langstrof, & Foulkes, 2013; Foulkes & Hay, 2015; Rácz, Hay, & Pierrehumbert, 2017; Weiss, Gerfen, & Mitchel, 2009), suggest that the two types of conditioning factors that make up sociolinguistic variables, social and linguistic, may interact more in processing than previously recognized. And, this interaction may have implications for not only social but also linguistic processing. Thus, listeners may have social motivations for monitoring probabilistic conditioning constraints on (ING) and other similar variables, and future work should test this hypothesis explicitly.

Conclusion

This study investigated the extent to which listeners track the systematicity of variability in speech, particularly the grammatical conditioning constraints of the sociolinguistic variable (ING). Our findings suggest that listeners can and do form probabilistic expectations about the realization of (ING), despite the fact that such expectations are not very informative for word recognition. Further, listeners rely more on these grammatical category expectations when the bottom-up cues in the signal are less reliable. These results have implications for theories of prediction in language processing, and bridge a gap between psycholinguistic and sociolinguistic accounts of probabilistic conditioning.

Acknowledgments

The authors wish to thank Matthew Bauer, Carissa Diantoro, Sam Elliott, Celina Maldonado, Omar Ortiz, Blair Prater, and Brandon Zuel for assistance with subject running and data coding. Thanks to our participants and stimuli producers for their time, and thanks also to audiences at several talks and conferences during the course of the project. We also thank Morgan Sonderegger, Volya Kapatsinski, Molly Babel, and the anonymous reviewers for their helpful feedback. All errors, of course, remain our own. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

A. Stimuli

See Table A1.

Table A1
Full list of experimental stimuli, by grammatical category of (ING) word.

Stimulus	(ING) word	Grammatical category of (ING) word
Anything good is fine with me	Anything	Pronoun3
I'd do *anything* to help him	Anything	Pronoun3
John would do *anything* for her	Anything	Pronoun3
Everything could go right and I'd still lose	Everything	Pronoun3
My boss said that *everything* is on the table	Everything	Pronoun3
It has *everything* to do with him	Everything	Pronoun3
I wore a *boring* outfit to that interview	Boring	Adjective
The *boring* class went on too long	Boring	Adjective
Mr. Jones was the most *boring* tutor	Boring	Adjective
You're just *charming* today, aren't you?	Charming	Adjective
The *charming* contestant was my favorite	Charming	Adjective
She's actually *charming* in person	Charming	Adjective
The *developing* nations are still behind	Developing	Adjective
The *emerging* attitude is a good one	Emerging	Adjective
The movie was really *exciting*	Exciting	Adjective
He told me a *frightening* story	Frightening	Adjective
This magazine is really *interesting*	Interesting	Adjective
That *interesting* guy is my favorite	Interesting	Adjective
That was an *interesting* discussion	Interesting	Adjective
The *playing* kid tripped and fell	Playing	Adjective
The *shopping* mall was crowded	Shopping	Adjective
Let *sleeping* dogs lie	Sleeping	Adjective
The *building* is on fire	Building	Noun
That story had an unexpected *ending*	Ending	Noun
Cheap *housing* is hard to come by in this town	Housing	Noun
The *morning* clears my mind	Morning	Noun
The other *morning* I overslept and missed the bus	Morning	Noun
Every *morning* Ted goes for a run	Morning	Noun
I go out for coffee every *morning*	Morning	Noun
We eat outside on sunny *mornings*	Mornings	Noun
The *warning* helped us prepare for the storm	Warning	Noun
He was the best at *climbing* the rock	Climbing	Gerund
I like *cooking* big breakfasts on Sundays	Cooking	Gerund
Stop *involving* me in your plans	Involving	Gerund
I realized that *reading* the textbook takes forever	Reading	Gerund
Last night I went *running* in the park	Running	Gerund
Shopping is a great idea	Shopping	Gerund
Sleeping does you wonders when you're sick	Sleeping	Gerund
Smoking outside is allowed	Smoking	Gerund
I prefer *swimming* in the ocean	Swimming	Gerund

(*continued on next page*)

Table A1 (*continued*)

Stimulus	(ING) word	Grammatical category of (ING) word
Thinking is hard when I first get up	Thinking	Gerund
I hate *washing* the dishes	Washing	Gerund
I'm *accounting* for it	Accounting	Progressive
The governor has been *admitting* guilt	Admitting	Progressive
This movie is *becoming* truly awesome	Becoming	Progressive
She's *becoming* the best candidate	Becoming	Progressive
She's *clamoring* to get in	Clamoring	Progressive
The bus was *coming* close to me	Coming	Progressive
I wish I knew when my roommate was *coming* home	Coming	Progressive
I'm *developing* a plan	Developing	Progressive
We're *developing* delta equations in class	Developing	Progressive
She's *emerging* as the leader in the race	Emerging	Progressive
I'm *feeling* great today	Feeling	Progressive
Detroit is *getting* cleaned out	Getting	Progressive
John's *heading* over later	Heading	Progressive
She's *heading* to the store	Heading	Progressive
My mom is finally *living* her childhood dreams	Living	Progressive
She's just *living* completely for free	Living	Progressive
He's *looking* to hit a homerun today	Looking	Progressive
The sheets are *looking* kind of ragged	Looking	Progressive
The kids are *looking* out for you	Looking	Progressive
He's *lying* about what happened	Lying	Progressive
She's been *lying* to me all week about it	Lying	Progressive
He's *making* country music good again!	Making	Progressive
It's *making* time for it that's hard	Making	Progressive
She's *paying* for her own education	Paying	Progressive
He's still *paying* to go to school	Paying	Progressive
I'm *paying* good money for this trip	Paying	Progressive
You're never *paying* attention to me	Paying	Progressive
The kids are *playing* together	Playing	Progressive
She's always *reading* kids books	Reading	Progressive
He's *running* crazy late today	Running	Progressive
I was *running* toward the tree when it fell over	Running	Progressive
She's *shopping* at the mall	Shopping	Progressive
He's *shopping* clear across town	Shopping	Progressive
He's *sitting* down on that chair	Sitting	Progressive
I'd hate to be *sitting* around all day at a desk	Sitting	Progressive
She's *sleeping* downstairs	Sleeping	Progressive
He's *struggling* to get better at chemistry	Struggling	Progressive
He's *struggling* in class this term	Struggling	Progressive
I was *swimming* quickly	Swimming	Progressive
He's *talking* crazy talk	Talking	Progressive
John was *talking* to his parents about his grades	Talking	Progressive
I'm just *thinking* out loud	Thinking	Progressive
She is *working* for the UO	Working	Progressive
She's *working* downtown	Working	Progressive
He's *working* continuously on his next album	Working	Progressive
Carla is *writing* a story about it	Writing	Progressive
I'm *writing* Carl a letter	Writing	Progressive
He's *writing* to his parents	Writing	Progressive
That test was so hard, *nothing* could help me	Nothing	Pronoun2
I've got *nothing* to say to her	Nothing	Pronoun2
That movie did *nothing* for me	Nothing	Pronoun2
Did you get me *something* I asked for?	Something	Pronoun2
He told me *something* completely wrong	Something	Pronoun2
I've got *something* to tell you	Something	Pronoun2

B. Stimuli (ING) acoustic measures and analysis

The (ING) words from the stimuli for Experiments 2 and 3 were subjected to acoustic measurement and analysis. Stimuli were first force-aligned at the phone level using the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), with a dictionary allowing both *–in* and *–ing* pronunciations as options. Hand-checking assured that phone alignments provided accurate boundaries for the vowel and nasal segments. Durations of the (ING) form and its vowel and nasal segments were extracted, and acoustic measurements were taken in Praat for the first, second, and third formant (F1, F2, and F3) of the vowel midpoint (indicators of vowel quality), for the bandwidth of the first and second formant (B1 and B2) of the vowel at midpoint (indicators of vowel nasality), and for the mean pitch and intensity of the entire (ING) (indicators of the prominence of the (ING)). Formant measures and pitch were transformed to the Bark psychoperceptual scale (Traunmüller, 1990). Tables B1 (Experiment 2) and B2 (Experiment 3) provide means for stimulus measurement values by *–in* and *–ing* realization, and the six grammatical

Table B1
Experiment 2 stimuli acoustic measures and analysis results.

	Realization	Grammatical category						Statistical significance		
		Pronoun-3	Adjective	Noun	Gerund	Progressive	Pronoun-2	Realization	Gram Cat	Realiz × Gram
(ING) duration (ms)	*–ing*	128.3	141.9	162.2	113.9	118.3	110.0	$p < 0.0001^*$	$p = 0.01^*$	$p = 0.15$
	–in	151.7	115.0	136.7	101.8	101.2	90.0			
(ING) nasal duration (ms)	*–ing*	83.3	97.5	98.9	65.5	71.0	61.7	$p = 0.009^*$	$p = 0.04^*$	$p = 0.10$
	–in	100.0	76.2	83.3	65.5	64.0	51.7			
(ING) vowel duration (ms)	*–ing*	45.0	44.4	63.3	48.2	47.3	48.3	$p < 0.0001^*$	$p = 0.002^*$	$p = 0.41$
	–in	51.7	38.8	53.3	36.4	37.3	38.3			
(ING) vowel F1 (Bark)	*–ing*	6.18	5.47	5.72	5.63	5.76	6.24	$p = 0.02^*$	$p = 0.45$	$p = 0.77$
	–in	6.40	5.92	6.12	5.82	5.93	6.10			
(ING) vowel F2 (Bark)	*–ing*	13.57	13.72	13.42	14.20	13.92	13.73	$p < 0.0001^*$	$p = 0.97$	$p = 0.35$
	–in	13.52	12.60	12.91	12.73	12.66	12.74			
(ING) vowel F3 (Bark)	*–ing*	15.86	15.73	15.95	16.42	15.92	16.27	$p = 0.14$	$p = 0.39$	–
	–in	16.12	15.39	15.69	16.15	15.81	16.18			
(ING) vowel B1 (Bark)	*–ing*	1.41	1.94	2.74	1.42	1.26	1.25	$p = 0.72$	$p = 0.22$	–
	–in	1.13	1.72	2.21	1.10	1.77	0.67			
(ING) vowel B2 (Bark)	*–ing*	5.08	3.10	3.04	2.41	3.96	3.49	$p = 0.28$	$p = 0.44$	–
	–in	1.81	3.34	2.27	2.82	3.65	1.60			
(ING) pitch (Bark)	*–ing*	2.30	2.21	2.16	2.44	2.30	2.19	$p = 0.08$	$p = 0.72$	–
	–in	1.96	2.21	2.29	2.45	2.16	2.28			
(ING) intensity (dB)	*–ing*	68.4	67.7	65.3	69.7	68.9	67.4	$p = 0.62$	$p = 0.44$	–
	–in	67.1	69.3	67.4	70.0	68.3	67.7			

Table B2
Experiment 3 stimuli acoustic measures and analysis results.

	Realization	Grammatical category						Statistical significance		
		Pronoun-3	Adjective	Noun	Gerund	Progressive	Pronoun-2	Realization	Gram Cat	Realiz × Gram
(ING) duration (ms)	*–ing*	135.0	142.5	165.6	112.7	120.4	113.3	$p = 0.0002^*$	$p = 0.001^*$	$p = 0.28$
	–in	158.3	118.8	142.2	99.1	101.0	91.7			
(ING) nasal duration (ms)	*–ing*	91.7	98.1	101.1	67.3	71.0	66.7	$p = 0.009^*$	$p = 0.004^*$	$p = 0.28$
	–in	106.7	77.5	87.8	61.8	64.0	56.7			
(ING) vowel duration (ms)	*–ing*	43.3	44.4	64.4	45.5	49.4	46.7	$p < 0.0001^*$	$p = 0.004^*$	$p = 0.20$
	–in	51.7	41.2	54.4	37.3	37.1	35.0			
(ING) vowel F1 (Bark)	*–ing*	6.12	5.81	6.00	5.53	5.86	6.28	$p = 0.0007^*$	$p = 0.26$	$p = 0.45$
	–in	6.45	6.20	5.92	5.93	6.10	6.25			
(ING) vowel F2 (Bark)	*–ing*	14.20	13.96	13.41	13.36	14.03	13.67	$p < 0.0001^*$	$p = 0.48$	$p = 0.75$
	–in	13.48	12.98	12.58	13.12	13.02	12.72			
(ING) vowel F3 (Bark)	*–ing*	16.16	15.65	15.83	15.71	15.92	15.98	$p = 0.13$	$p = 0.67$	–
	–in	16.11	15.76	15.77	16.26	16.08	15.81			
(ING) vowel B1 (Bark)	*–ing*	1.34	1.84	2.09	1.23	1.31	1.09	$p = 0.72$	$p = 0.01^*$	$p = 0.82$
	–in	1.24	1.46	2.44	1.19	1.34	0.66			
(ING) vowel B2 (Bark)	*–ing*	2.61	3.41	4.00	4.67	4.37	2.97	$p = 0.002^*$	$p = 0.13$	$p = 0.96$
	–in	1.70	2.09	2.13	2.48	3.32	1.71			
(ING) pitch (Bark)	*–ing*	2.29	2.22	2.08	2.47	2.30	2.18	$p = 0.08$	$p = 0.57$	–
	–in	1.92	2.21	2.28	2.45	2.15	2.27			
(ING) intensity (dB)	*–ing*	68.2	68.2	65.5	69.3	69.2	67.6	$p = 0.11$	$p = 0.22$	–
	–ing	66.0	69.2	66.3	69.6	67.9	67.7			

categories. The tables also provide summaries of statistical significance for the two factors of interest: speakers' realization of (ING) as *–in* or *–ing* and the grammatical category of the (ING) word, as main effects and, where warranted, in interaction. Mixed-effect linear regressions were built and compared via likelihood ratio tests. All models contained random intercepts for speaker and for item (stimulus sentence). Statistical significance (indicated by ∗ for $p < 0.05$) for realization and for grammatical category, as main effects, was assessed and the *p* value provided in the tables is based on a likelihood ratio test comparing a model with each main effect to an intercept-only model. For cases where at least one of the main effects was significant, significance for the interaction is based on a model with the interaction between realization and grammatical category and the main effects tested against a model with just main effects for the two factors.

References

Abramowicz, Ł. (2007). Sociolinguistics meets exemplar theory: Frequency and recency effects in (ing). University of Pennsylvania working papers in linguistics, 13(2), 3.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 115*(6), 3171–3183.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech, 56*(3), 329–347.

Bender, E. M. (2005). On the boundaries of linguistic competence: Matched-guise experiments as evidence of knowledge of grammar. *Lingua, 115*(11), 1579–1598.

Boersma, P., & Weenink, D. (2015). Praat version 5.4. 08. Doing phonetics by computer.

Bresnahan, M. J., Ohashi, R., Nebashi, R., Liu, W. Y., & Shearman, S. M. (2002). Attitudinal and affective response toward accented English. *Language & Communication, 22*(2), 171–185.

Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language, 86*(1), 168–213.

Brouwer, S., Mitterer, H., & Huettig, F. (2013). Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics, 34*(3), 519–539.

Bürki, A., Ernestus, M., & Frauenfelder, U. H. (2010). Is there only one "fenêtre" in the production lexicon? On-line evidence on the nature of phonological representations of pronunciation variants for French schwa words. *Journal of Memory and Language, 62*(4), 421–437.

Campbell-Kibler, K. (2007). Accent, (ING), and the social logic of listener perceptions. *American Speech, 82*(1), 32–64.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

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

Connine, C. M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin & Review, 11*(6), 1084–1089.

Connine, C. M., & Clifton, C., Jr (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 13*(2), 291.

Connine, C. M., Ranbom, L. J., & Patterson, D. J. (2008). Processing variant forms in spoken word recognition: The role of variant frequency. *Attention, Perception, & Psychophysics, 70*(3), 403–411.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words.* Cambridge, MA: The MIT Press.

Davies, M. (2008). *The corpus of contemporary American English.* BYE: Brigham Young University.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research, 229*(1), 132–147.

Docherty, G. J., Langstrof, C., & Foulkes, P. (2013). Listener evaluation of sociophonetic variability: Probing constraints and capabilities. *Linguistics, 51*(2), 355–380.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One, 8*(10), e77661.

Fischer, J. L. (1958). Social influences on the choice of a linguistic variant. *Word-Journal of the International Linguistic Association, 14*, 47–56.

Forrest, J. (2015). Community rules and speaker behavior: Individual adherence to group constraints on (ING). *Language Variation and Change, 27*(3), 377–406.

Foulkes, P., & Hay, J. B. (2015). The emergence of sociophonetic structure. *The Handbook of Language Emergence,* 292–313.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

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

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*(1), 58–93.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review, 11*(4), 716–722.

Hazen, K. (2008). (ING): A vernacular baseline for English in Appalachia. *American Speech, 83*, 116–140.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*(3), 346–363.

Houston, A. (1985). Continuity and change in English morphology: The variable (ING). Unpublished doctoral dissertation, University of Pennsylvania.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118–135.

Hura, S. L., Lindblom, B., & Diehl, R. L. (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech, 35*(1–2), 59–72.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. Mullennix (Eds.). *Talker variability in speech processing* (pp. 145–166). San Diego: Academic Press.

Kendall, T. (2008). Accommodating (ING): Individual variation in mixed-ethnicity interviews. In B. Heselwood, & C. Upton (Eds.). *Proceedings of methods XIII: Papers from the thirteenth international conference on methods in dialectology* (pp. 351–361). Frankfurt am Main: Peter Lang.

Kendall, T. (2013). *Speech rate, pause and sociolinguistic variation: Studies in corpus sociophonetics.* New York: Palgrave Macmillan.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2014). LmerTest: Tests for random and fixed effects for linear mixed effect models. R package, version 2.0-3.

Labov, W. (1966). *The social stratification of English in New York City.* Washington: Center for Applied Linguistics.

Labov, W. (2001). *Principles of linguistic change.* Oxford: Blackwell.

Labov, W., Ash, S., Ravindranath, M., Weldon, T., Baranowski, M., & Nagy, N. (2011). Properties of the sociolinguistic monitor. *Journal of Sociolinguistics, 15*(4), 431–463.

Levon, E., & Fox, S. (2014). Social salience and the sociolinguistic monitor: A case study of ing and th-fronting in Britain. *Journal of English Linguistics, 42*(3), 185–217.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.

Loudermilk, B. C. (2013). Cognitive mechanisms in the perception of sociolinguistic variation. Unpublished doctoral dissertation, University of California, Davis.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics, 39*(3), 155–158.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. *Handbook of psycholinguistics* (pp. 581–611). (2nd ed.). .

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human perception and performance, 15*(3), 576.

McAuliffe, M., & Babel, M. (2016). Stimulus-directed attention attenuates lexically-guided perceptual learning. *Journal of the Acoustical Society of America, 140*(3), 1727–1738.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In Proceedings of Interspeech.

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*, 219–246.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America, 131*(1), 509–517.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory phonology VII* (pp. 101–139). Berlin: Mouton de Gruyter.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics, 4*(34), 516–530.

Pinnow, E., & Connine, C. M. (2014). Phonological variant recognition: Representations and rules. *Language and Speech, 57*(1), 42–67.

Pitt, M. A. (2009). How are pronunciation variants of spoken words recognized? A test of generalization to newly learned words. *Journal of Memory and Language, 61*(1), 19–36.

Pitt, M. A., Dilley, L., & Tat, M. (2011). Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics, 39*(3), 304–311.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance, 19*(4), 699.

Preston, D. R. (2013). Language with an attitude. In J. K. Chambers, & N. Schilling-Estes (Eds.). *The handbook of language variation and change* (pp. 157–182). (2nd ed.). Malden, MA: Wiley-Blackwell.

Psychology Software Tools, Inc. (2012). E-Prime 2.0 [computer software].

Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology, 8*, 51.

Ranbom, L. J., & Connine, C. M. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language, 57*(2), 273–298.

Sumner, M. (2013). A phonetic explanation of pronunciation variant effects. *The Journal of the Acoustical Society of America, 134*(1), EL26-EL32.

Sumner, M., Kurumada, C., Gafter, R., & Casillas, M. (2013). Phonetic variation and the recognition of words with pronunciation variants. In The 35th annual meeting of the cognitive science society (pp. 3486–3492). Cognitive Science Society.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4*, 1–13.

Tagliamonte, S. (2004). Somethi[ŋ]'s goi[n] on!: Variable (ing) at ground zero. In B.-L. Gunnarsson, L. Bergström, G. Eklund, S. Fidell, L. H. Hansen, A. Karstadt, & M. Thelander (Eds.). *Language variation in Europe: Papers from the second international conference on language variation in Europe, ICLaVE 2* (pp. 390–403). Uppsala, Sweden: Uppsala Universitet.

Tamminga, M. (2017). Matched guise effects can be robust to speech style. *The Journal of the Acoustical Society of America, 142*(1), EL18-EL23.

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, & Psychophysics, 77*(5), 1674–1684.

Theodore, R. M., Myers, E. B., & Lomibao, J. A. (2015). Talker-specific influences on phonetic category structure. *The Journal of the Acoustical Society of America, 138*(2), 1068–1078.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America, 88*(1), 97–100.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*(3), 528–553.

Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*, 580–591.

Vaughn, C. & Kendall, T. (2018). Stylistic covariation and variable (ING). Paper presented at LabPhon 16. Lisbon, Portugal, June 2018.

Wald, B., & Shopen, T. (1981). A researcher's guide to the sociolinguistic variable (ING). In T. Shopen, & J. M. Williams (Eds.). *Style and variables in English* (pp. 219–249). Winthrop Publishers.

Walker, A., & Hay, J. B. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology, 2*, 219–237.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*(3917), 392–393.

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? *Language Learning and Development, 5*(1), 30–49.

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *Journal of Neuroscience, 32*(40), 14010–14021.

Zarcone, A., Van Schijndel, M., Vogels, J., & Demberg, V. (2016). Salience and attention in surprisal-based accounts of language processing. *Frontiers in Psychology, 7*(844), 1–17.