Experience with a linguistic variant affects the acquisition of its sociolinguistic meaning: An alien-language-learning experiment

Wei Lai[1], Péter Rácz[2], and Gareth Roberts*[1]

[1]Department of Linguistics, University of Pennsylvania
[2]Cognitive Development Center, Central European University
Affiliation

Abstract

How do speakers learn the social meaning of different linguistic variants, and what factors influence how likely a particular social–linguistic association is to be learned? It has been argued that the social meaning of more salient variants should be learned faster, and that learners' pre-existing experience of a variant will influence its salience. In this paper we report two artificial-language-learning experiments investigating this. Each experiment involved two language learning stages followed by a test. The first stage introduced the artificial language and trained participants in it, while the second stage added a simple social context using images of cartoon aliens. The first learning stage was intended to establish participants' experience with the artificial language in general and with the distribution of linguistic variants in particular. The second stage, in which linguistic stimuli were accompanied by images of particular aliens, was intended to simulate the acquisition of linguistic variants in a social context. In our first experiment we manipulated whether a particular linguistic variant, associated with one species of alien in the second learning phase, had been encountered in the first learning phase. In the second experiment we manipulated whether the variant had been encountered in the same grammatical context. In both cases we predicted that the unexpectedness of a new variant or a new grammatical context for an old variant would increase the variant's salience and facilitate the learning of its social meaning. This is what we found, although in the second experiment, the effect was driven by better learners. Our results suggest that unexpectedness increases the salience of variants and makes their social distribution easier to learn, deepening our understanding of the role of individual language experience in the acquisition of sociolinguistic meaning.

Keywords: artificial-language learning; social meaning; sociolinguistics; salience; surprisal; expectation violation

*Correspondence should be sent to Gareth Roberts, Department of Linguistics, University of Pennsylvania, 3401-C Walnut Street, Suite 300, Philadelphia, PA, 19104, USA. E-mail: gareth.roberts@ling.upenn.edu

Introduction

Language varies, and linguistic variants convey social identity, both in the sense that language users make social inferences about people based on characteristics of their language (Labov et al., 2011) and in the sense that they make linguistic inferences about people based on the social groups they belong to (Bell, 2001). In this paper we report two artificial-language-learning experiments that investigate how language users acquire associations between a social group and a linguistic variant used by its members, and how this acquisition is modulated by the learner's prior experience with the variant in question.

Acquisition of the social meaning of linguistic variants

Language exhibits variation at every level of organization. That is, speakers vary in the lexical items they use to refer to things (e.g., *pavement* vs. *sidewalk*), in the phonemes these items are constructed from (e.g., whether the vowel in *pin* is distinct from the vowel in *pen*), in the morphosyntactic rules for combining them (e.g., whether "the band are playing" is grammatical), and so on. Because the distribution of variation is related in large part to the distribution of speakers, particular linguistic variants[1] can serve as identity markers by signaling users' membership in certain social categories. Previous studies have reported that listeners exploit linguistic features to make social inferences about speakers, including such characteristics as ethnicity (Purnell, Idsardi, & Baugh, 1999), regional origin (Clopper & Pisoni, 2004), sexuality (Munson, McDonald, DeBoe, & White, 2006), and personality characteristics (e.g., Campbell-Kibler, 2007; Lambert, 1967). Speakers also converge towards or diverge from their interlocutors' speech depending on their social evaluations of their interlocutors (e.g., Babel, 2010, 2012; Beňuš, 2014). Moreover, a growing body of experiments on speech perception and processing has shown that listeners take into account perceived social identity in interpreting speech. For example, listeners adjust phonemic boundaries depending on speakers' perceived gender (Strand, 1999), age (Hay, Warren, & Drager, 2006), and nationality (Niedzielski, 1999), leading the same acoustic signal to be interpreted as different lexical items. Taken together, these findings suggest that, far from being incidental to language, socioindexical meaning is tightly interwoven with non-social linguistic knowledge.

How then are associations between linguistic variants and social identity acquired? This question was first addressed in research examining how adult linguistic variants are distributed in child language production (e.g., Labov, 1989; J. Roberts & Labov, 1995; Smith, Durham, & Fortune, 2009). These studies found evidence that children match the variant distribution of their parents as early as 3–4 years old, and some dialect-specific constraints are acquired before non-dialect-specific constraints (Labov, 1989). The gender stratification of variation in adult speech has also been found in child speech: Foulkes, Docherty, and Watt (2005) examined whether the gender stratification in preaspiration before /t/ in Newcastle-upon-Tyne (Docherty & Foulkes, 2005) was mirrored in child speech, and found

[1] There is reason to say that it is variants and not variables that serve as identity markers. While pronouncing English *-ing* as *-in* will have a particular social meaning in a particular context, for instance, it is not the case that pronouncing it as *-ing* will necessarily have the inverse meaning. Different variants of the same variable, in other words, can have independent social meanings from each other (Campbell-Kibler, 2010).

different preaspiration patterns between boys and girls among children aged 3;6–4;0, but not among those aged 2;0–3;6.

In comparison, the question of how adults come to learn the association between linguistic variants and social categories is relatively understudied, partly due to the complexity of the multiple potential associations between linguistic cues and indexical meanings in the real world, as well as the difficulty in quantifying individual adults' experience with them. Foulkes, Docherty, Khattab, and Yaeger-Dror (2010), for example, investigated whether adults from Newcastle are more likely than adults from elsewhere to judge children's speech samples containing preaspirated /t/ as having been produced by girls. They found, however, that the additional sensitivity to preaspirated tokens as an index of females was confounded by other gender-differential acoustic parameters such as loudness, F_0 and speech rate.

To assist in ruling out these kinds of confounds, researchers have begun to complement natural-language research with work investigating the acquisition of associations between artificial linguistic variants and fictional social categories. One of the earliest studies along these lines was that of Docherty, Langstrof, and Foulkes (2013), who evaluated the acquisition of associations between realistic phonetic variants and novel social-category labels through passive exposure to word-label pairs. They found that participants were able to learn socially structured phonetic variation with a relatively small amount of exposure. However, these experiments were still not independent of real-world influences, as they used natural-language stimuli with existing (and, to the participants, already familiar) phonetic variables, such as [bʌtə] and [bʌʔə] for "butter". As the authors noted, this led to their results being skewed by the nature of their stimuli and language experience of their participants.

Rácz, Hay, and Pierrehumbert (2017) also investigated the acquisition of novel social–linguistic associations but used an entirely artificial language. They trained participants by having them play a game involving a series of forced-choice questions, and found that adult players learned an alternation between morphological variants dependent on the gender of the interlocutor, and were able to generalize the association to novel lexical items in similar contexts. Interestingly, Rácz et al. (2017) found that, although the association between morphological variants and interlocutor gender was learnable, an association between the same variants and the spatial orientation of the interlocutor was barely acquired. There are various ways in which gender and spatial orientation differ from each other, with the latter being for instance less permanent and more contingent than the former. However, Rácz et al. (2017) emphasized the role of social *salience*: While both contexts used in the Rácz study were visually distinct, only gender has any kind of prominence as a cue of social language use in the participants' prior experience. This finding resonates with other research.

The salience and learnability of sociolinguistic variants

As pointed out by Jaeger and Weatherholtz (2016), the term "salience" is rarely defined very precisely in the sociolinguistic literature and turns out to be rather hard to define precisely. For present purposes we will define it somewhat broadly in terms of relative prominence; to put it another way, something that is relatively salient is something that is relatively likely to attract attention. This property has been invoked in a great deal of classic and recent sociolinguistic literature. Labov (1972), for instance, divided sociolinguistic variants into indicators, markers and stereotypes based in large part on their level of salience, as reflected in their varying propensity to be involved in social stratification, style shifting, explicit meta-commentary, hypercorrection and stigmatization. Consequently, variants with higher salience are encoded with more attention and higher meta-linguistic awareness, leading them to be more easily recognized and retained in memory than other variants with equal frequency (Sumner, Kim, King, & McGowan, 2014), resulting in a potential acquisition bias that cannot be explained by frequency of exposure alone (Jaeger & Weatherholtz, 2016).

The findings of other artificial-language studies on sociolinguistic questions (e.g., G. Roberts, 2010; Sneller & Roberts, 2018) are also highly relevant to salience, though their authors do not always discuss it explicitly. Sneller and Roberts (2018), for example, investigated the spread of dialectal variants between different "alien" groups in a communicative trading game, finding that the spread of a variant was influenced both by the kind of socio-indexical meaning it had (i.e., whether it indexed a group or a perceived trait of that group) and the practical relevance of that meaning in the context of the game. As with Rácz et al. (2017), these findings can be framed in terms of social salience. Broadly speaking, work in a range of paradigms suggests that the probability with which speakers perceive and adopt novel sociolinguistic variants is affected by the salience of those variants, which is not constant and varies with context.

In fact there are quite a number of factors that have been argued to affect the salience of a linguistic variant, including: (a) *frequency of exposure*, with variants occurring at higher frequency typically being more salient (cf. Labov et al., 2011); (b) *physical properties* intrinsic to the variant, such that variants with higher phonetic discreteness and prosodic prominence are more salient (cf. Kerswill & Williams, 2002); (c) the *social context* in which the variant is embedded; for example, as discussed above, the salience of a variable may vary depending on whether the relevant social factor indexed is sufficiently evident to learners (e.g., Rácz et al., 2017), or whether the social pressure is strong enough to motivate active use of linguistic variables as social markers (e.g., Sneller & Roberts, 2018); (d) the individual's pre-existing nonlinguistic experience, beliefs or ideology, which might skew the statistical properties of a variant in their mind through mediation of attentional and attitudinal factors (Levon, 2006); and (e) the individual's pre-existing linguistic experience of the variant, including whether or not it has been encountered before, and – if so – what constrained it. In the study presented here, we focus on the last of these factors, which will be discussed in the next section.

Salience emerging from the violation of linguistic experience

The present paper focuses on the influence of individual experiential factors on sociolinguistic acquisition. Previous work in this vein has paid much attention to *non-linguistic* experience such as social and developmental experience (e.g., Foulkes & Docherty, 2006) or social stereotypes (Levon, 2006). By comparison, the influence of *linguistic* experience on sociolinguistic acquisition is relatively understudied. In particular, there is very little work on how the salience of a sociolinguistic variant is affected by prior experience of that variant in other contexts. Theoretically, Jaeger and Weatherholtz (2016) proposed that salience related to language experience can be better understood "with reference to prior expectations based on listeners' past language experience and the ensuing expectation violation that a listener experiences relative to those prior expectations" (p. 1). One analogue they provided was the well-attested novelty bias effect; that is, novel items and surprising events that we do not expect tend to stand out, gaining more attention and stronger encoding. They argued that a similar novelty bias might occur for a first-encountered linguistic variant: When a lect is encountered for the first time, listeners may find that some feature violates their expectations owing to its absence in their prior language experience; the expectation violation causes a level of surprisal/unexpectedness and contributes to the perceptual salience of this variant. A strength of defining salience as a function of surprisal or violation of expectations is that it allows the quantification of cognitive consequences, since the degree to which something is violating our expectations is measurable in an information-theoretic framework (Hale, 2001; Levy, 2008; Shannon, 1948), according to which the surprisal/unexpectedness associated with an input can be quantified in terms of the amount of new information gained by processing the input (see Jaeger & Weatherholtz, 2016, for more in-depth elaboration).

Although Jaeger and Weatherholtz's (2016) approach seems appealing for its provision of operational ways to quantify salience emerging from expectation violation, neither it nor its potential benefit for acquisition has, to the best of our knowledge, received much experimental attention. There are a number of experimental studies on language processing showing that less expected words and structures take longer, and are more costly, to process (e.g., McDonald & Shillcock, 2003; McRae, Spivey-Knowlton, & Tanenhaus, 1998), and similar processing costs have been observed for unfamiliar lectal variants encountered in comprehension tasks (Fraundorf & Jaeger, 2016; Kaschak & Glenberg, 2004; Squires, 2014). However, additional processing effort for novel variants in comprehension does not necessarily result in better performance in noticing or memorizing these variants or in associating them with the correct social group.

The present study is designed to make progress in pursuing the hypothesis that experience-dependent salience can arise from expectation violation, and (crucially) cause a sociolinguistic variant to be more learnable. We used an "alien language" learning paradigm in which participants first learned a miniature language and were then exposed to it in a social context with alien "interlocutors". Two types of expectation violation were simulated. The first involved a relatively strong violation, namely exposure to a variant that had been entirely absent in the learner's prior language experience. The second involved an intuitively subtler violation: exposure to a variant that was already familiar from prior experience, but which now occurred in a new grammatical context. If expectation violation contributes additional salience to sociolinguistic variants and if salience aids the acquisition of sociolinguistic associations, then we should expect learning performance to improve when the sociolinguistic variants to be learned violate prior language experience than when they conform to it.

Naturally, the idea that more attention is given to the unexpected is widely invoked in accounts of cognitive processing, from low-level systems like visual attention (Itti & Koch, 2000) and object perception (Hespos & Baillargeon, 2008) to complex reasoning (Stahl & Feigenson, 2015). The same is true for language, where salience and surprisal have been discussed in the context of sentence comprehension (Levy, 2008) and higher-level systems (Zarcone, Van Schijndel, Vogels, & Demberg, 2016). In sociolinguistics, they are relevant to questions of how variants propagate, and to what degree this process involves selection on the basis of their social meaning or neutral evolution, in which selection is absent and variants propagate through sampling error, mediated by variation in the frequency of interaction (Harrington, Kleber, Reubold, Schiel, and Stevens 2018; Trudgill 2008; Labov

2001, pp. 191, 506; see G. Roberts 2013, pp. 626–628 for a summary of this debate in the context of the emergence of New Zealand English). In cases where selection is assumed to occur, it is widely recognized that social factors may play an important role, but it remains an open question why certain variants attract particular social meanings while others don't. Salience, in some sense, seems an obvious candidate but the details of what might make a variant more salient than others have been little explored. Our experiments are designed to break ground on that question.

Artificial language studies of the sort we present here have tended, for simplicity's sake, to assume categorical associations between variants and social groups (e.g., Rácz et al., 2017; Sneller & Roberts, 2018). This is somewhat surprising, given that artificial-language studies of language acquisition, as opposed to sociolinguistic acquisition, have made progress in investigating the role of variation in input (e.g., Culbertson, Smolensky, & Legendre, 2012; Hudson Kam & Newport, 2009). Furthermore, the social distribution of linguistic variants is very often not categorical in the real world, something which Docherty et al. (2013) took into account in their study, in which particular variants were used by social groups only 80% of the time. We followed them in the study presented here in employing a non-categorical social distribution of variants in our artificial language.

Main questions and hypotheses

The main question of our study is as follows. The violation of expectation has been claimed to make a variant more salient (Jaeger & Weatherholtz, 2016). If this is so, will people better learn to associate a linguistic variant with the social group that uses it if the variant violates their expectations? We predicted that they would, and we investigated two kinds of expectation violation, hypothesizing that participants would be more likely to learn a social–linguistic association if (Experiment 1) they had not encountered it before and (Experiment 2) they had encountered it before, but subject to grammatical constraints that now appear to be violated. As an example of the first kind of violation, we might imagine an American English speaker visiting the Merseyside area of England and hearing, for the first time, *book* pronounced with a final velar fricative [x] (as in German *Buch*) instead of the expected velar stop [k]. As an example of the second kind of violation, consider a speaker who has heard *-th* pronounced as [f], but only at the end of syllables (as in [boʊf] for *both* or [ˈɛfnɪk] for *ethnic*). For a speaker who had acquired syllable-finality as a constraint on this variant, hearing [fɪŋk] for *think* would likely be at least somewhat salient. We also predicted that, in both experiments, participants who learned the sociolinguistic association would extend it to new items.

As discussed above, we tested our hypotheses in an artificial language experiment in which the association between variants and social groups was non-categorical. This has advantages in terms of ecological validity, but it also has the consequence that it matters how one asks if the target association has been acquired by participants. If we consider the example of velar fricatives in Merseyside English, we could imagine, on the one hand, asking other English people which speakers might pronounce "book" as [bʊx]. On the other hand, we could ask people how a Merseyside speaker would say "book". Importantly, the two questions could very well yield different impressions of the association between velar fricatives and social groups in England. This is in part due to likely differences in how people approach the question, but also to the non-categorical nature of the relationship.

The pronunciation [bʊx] is not common outside Merseyside, but that does not mean that Merseyside speakers rarely pronounce it [bʊk]. We took this asymmetry into account in our experiment and asked the question in both ways, predicting that responses would point in the same direction, but were likely to vary in how reliably they do so.

In summary, we predicted that participants would be able to learn to make an association between a linguistic variant and a social group in an artificial language, in which the variant and the social group did not co-occur categorically, and extend it to new items. That is, if exposed to two variants and two social groups, where each variant is more a feature of one group's language than the other, participants should learn that relationship and – when asked – associate the "right" variant with the right group more than the wrong variant. We also predicted that they would learn the association better if it violated expectations (either by being entirely novel or by occurring in a novel grammatical environment); this should show up in terms of a larger gap between the association of the "right" variant with the right group and the "wrong" variant with the same group for participants whose expectations had been violated. Finally we predicted that we should see the same effect whether we elicited a social group based on a variant or a variant based on a social group, but that the strength of the association would be likely to vary.

Experiment 1: First Encounter

Experiment Overview

We used an artificial-language-learning paradigm to investigate whether encountering a linguistic variant for the first time in a new social context would facilitate learning the linguistic variant as a social marker of that group. Participants were trained on an alien language with two dialects, each used by a different alien species, the *Gulus* or the *Norls*. The dialects differed with regard to a plural suffix: Gulus used *-dup* as the only form of the plural suffix whereas Norls sometimes used *-dup* but mostly used *-nup*. Sociolinguistic acquisition was evaluated on the basis of whether participants could infer which alien might have used a given suffix and, conversely, which suffix a given alien might have used. The training process consisted of two stages: Participants were first trained on the language without seeing any aliens, which was intended to establish "prior experience" with the language; they were then exposed further to the language, this time with alien interlocutors, which allowed them to learn associations between plural suffixes and alien species. Crucially, we manipulated participants' prior experience with the variant *-nup* such that half the participants would only ever encounter *-dup* in the first stage (*NoExposure* condition) whereas the other half would see both suffixes (*Exposure* condition). We predicted that participants with no experience of *-nup* in the first training phase would find it more salient in the second phase and better learn to associate it with Norls.

Method

Participants. 100 participants completed Experiment 1 online within the maximum allowed time (1.5 hours).[2] To help remove problematic participants who were (e.g.) not

[2] Across experiments 1 and 2, a further 10 participants took over an hour and a half, and their data were automatically excluded. We paid them for their participation nonetheless.

paying attention to the task, simply clicking buttons randomly, or taking lengthy breaks, we excluded any participants whose participation duration was below the 2.5% quantile or above the 97.5% quantile of all participants. (This was a preplanned exclusion conducted before analysing the data.) After exclusions we were left with 93 participants, 51 female and 42 male, aged 17–73 (mean: 29.02). Of these, 29 were recruited from the University of Pennsylvania Psychology subject pool and 64 were recruited through the Prolific Academic website. 49 participants were in the *Exposure* condition and 44 in the *NoExposure* condition.

Alien language. The "alien" language in training was composed of 14 word stems, as shown in Table 1, and two variants of a plural suffix: *-dup*, *-nup*. The 14 stem words were randomly generated by combining one or two CV syllables with a word-final CVC syllable from a segment pool of five vowels /a e i o u/ and 12 consonants /k g q h m n t s z j l w/.

Table 1
Stem Words in the Alien Language

nesel	laniz	wumos	maqub	firot	tugan	wukin
jemulok	nuwik	sehilod	gequzis	takoles	falon	hiwen

Aliens. The language was used by two alien species: ***Gulus*** and ***Norls***. The two species showed salient differences in their appearance: Gulus had green skin, a slim body and four arms while Norls had larger bodies, red skin and three eyes (see Fig. 1 for examples). Six idiosyncratic aliens were designed within each species by introducing variability in their attire and facial features. This was intended to ensure that the linguistic variation on the group level would not be mistaken for variation on the individual level.

The two dialects used by Norls and Gulus shared the same stem forms but exhibited different distributions of suffix variants: Gulus attached *-dup* to all 14 words to signal plurality, whereas Norls attached *-dup* to only four of the words (*hiwen*, *wukin*, *jemulok* and *wumos*) and *-nup* to the remaining ten words. Put differently, Gulus used the *-dup* variant 100% of the time as their plural suffix whereas Norls used *-dup* and *-nup* at a ratio of 29% to 71%.

Procedure. Fig. 1 displays the structure of Experiment 1. It involved two learning phases and a test phase. In the first learning phase, participants were trained on the alien words through passive exposure to word-object pairs and multiple-choice exercises with feedback. Objects associated with plural words were simple tripled versions of the objects associated with singular words.[3] During this phase, images of aliens were absent. After the first learning phase had finished, however, the two alien species were introduced to participants. Then, in the second learning phase, participants were presented with the same word-object pairs again, together with alien interlocutors. In the test phase we asked two types of questions. One involved presenting participants with an alien and asking which plural form they would expect the alien to produce; the other involved presenting participants with a plural form and asking which species of alien most likely produced it. Responses to these two types of questions formed our main dependent variable.

In what follows, we walk the reader through each phase in detail.

First learning phase: Learning without aliens. The experiment started with a learning phase that exposed participants to the words of the language without any aliens

[3] The images of the objects were created by Kayo Takasugi and are copyrighted to Northwestern University.

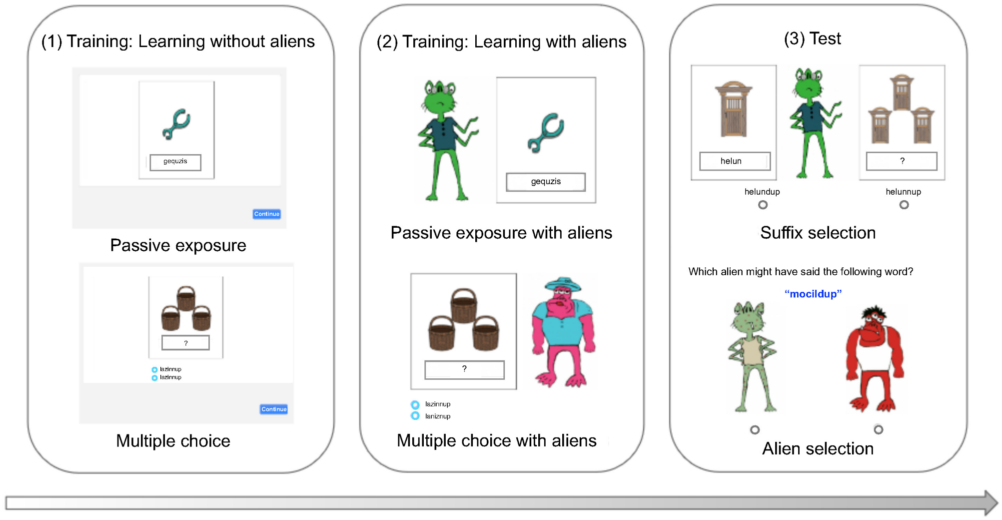

Figure 1. Structure of Experiment 1

present. This was designed to establish experience of the language before introducing it in a social context; by varying the contrast between prior experience and the following social context we could then manipulate the extent to which participants' expectations were violated. (We did not include aliens in the first learning phase, as this could introduce extraneous or nuisance variables by adding an extra social contrast on top of the linguistic contrast.) Panel 1 of Fig. 1 shows the two kinds of trials in this learning phase. On the top is a passive exposure trial in which participants were presented with a word paired with an image of the object(s) it referred to. Participants were instructed to memorize the word and its meaning before clicking on a CONTINUE button to proceed to the next trial. In a forced-choice trial, as shown in the bottom of Panel 1, participants were presented with an image and were asked to choose the correct word for it from two options, one correct and one a foil that was not a word in the language. The foil was generated by changing one or two segments of the correct form on the stem but not the suffix (e.g., *lanizdup* vs. *lazindup*). During the learning phase participants were exposed to 28 alien words (14 singular, 14 plural) in total. These were divided into seven sets of four words each, with the constraint that each set contained two singular words and two plural words that all had different stems from each other. For each word set, participants would go through passive exposure to each of the four word stimuli, and then complete two-way forced-choice exercises on them. For the forced-choice trials, participants would receive one point for each correct response and no point for an incorrect one. (Over the course of the whole experiment, a participant could gain a maximum of 168 points.) Feedback on the correct form was provided immediately afterwards, as was the point for a correct response. Participants had to complete training on one word set before proceeding to the next. The order of the seven word sets was randomized for each participant, as was the order of the four trials within each passive exposure section and within each forced-choice section. The whole process was repeated once participants had completed training on all seven word sets. In total, participants were exposed to 14 words × 2 forms (singular and plural) × 2 training (exposure and exercise) × 2 repetitions = 112 trials in this stage.

Second learning phase: Learning with aliens. In the interval between the first and the second training stages, there was a short familiarization phase during which the aliens were introduced. Participants were first presented with images of Gulus and Norls, each labeled with the species name; then the labels were removed and participants were instructed to drag and drop each alien into one of the the two boxes labeled Gulu and Norl. Feedback was provided after the drag-and-drop. After familiarization, the second learning stage started, in which participants saw not only object-word pairs but also an alien "interlocutor", as shown in Panel 2 of Fig. 1. The learning process resembled the previous stage in its structure, except that each word set contained eight words, four from Gulus and four from Norls. Words within each set still conformed to the constraint that they were composed of an equal number of plural forms and singular forms and could not have the same stem. In total, each participant was exposed to 14 words × 2 forms (singular, plural) × 2 species (Gulu, Norl) × 2 training (exposure, exercise) × 2 repetitions = 224 trials.

Test: Sociolinguistic acquisition. At the end of the learning phase, participants proceeded to the test phase, which evaluated the extent to which participants had established associations between alien groups and plural suffixes. As explained above, there are two ways to make this evaluation, and we employed both so that we could compare the two. This meant that the test phase contained both a suffix-selection task – in which participants were presented with an alien from one species and had to select which plural suffix the alien might use – and an alien-selection task, in which they were presented with a plural word and had to select which alien was most likely to have said it. Trials in the Test phase contained both *old word stimuli* from the learning phase and *new word stimuli* that participants had never seen. Through the inclusion of novel stems, both tasks thus evaluated whether participants' behavior could be generalized to new lexical items that did not appear in the training phase. Questions were presented in a random order for each participant, and – within each question – the on-screen order of the two possible answers was always counterbalanced. No feedback was provided afterwards.

In **suffix selection**, trials on *old words* worked like forced-choice trials in the second learning phase (i.e., as in the bottom of Panel 2 in Fig. 1), except that the stems of the optional answers were identical and the suffixes different. Participants were instructed to choose the form the alien interlocutor would likely use. Trials on the *new words* were different: As shown on the top of Panel 3, participants were presented with a singular word and an image of the object it referred to, as well as an alien interlocutor; they were required to choose the plural form, with a choice between a *dup*-ending word and a *nup*-ending word. In all, the task included 56 trials on old words (14 words × 2 species × 2 repetitions = 56 trials), 20 trials on new words (5 new words × 2 species × 2 repetitions = 20 trials), and 34 filler trials which tested on either singular words or plural words with incorrect stems.

In **alien selection**, participants were given a plural word and had to choose between a Gulu and a Norl as the likely speaker of the word, as in the bottom of Panel 3 in Fig. 1. The same two individual aliens were used throughout the whole task, to avoid the possibility that participants might associate different members of the same species with different forms, but the species order was counterbalanced across questions. The prompt words were generated by attaching the 14 learned words and the five new words once each with *-dup* and once each with *-nup*, such that there were 38 trials (14 old word × 2 suffixes + 5 new word × 2 suffixes) in total.

We kept a tally of each participant's test score, defined as the number of correct answers the participant gave in test. This score was not shown to the participant.

Between-participant conditions. Participants were randomly assigned to two between-participant conditions: the NoExposure condition and the Exposure condition. The two conditions differed with respect to the presence or absence of *-nup* in the first learning stage: Participants in the NoExposure condition would see all the plural words exclusively affixed by *-dup*, whereas participants in the Exposure condition would see four *nup*-ending plural forms (29%) and 10 *dup*-ending plural forms (71%) in the first stage. The two conditions were identical in the second learning stage: Gulus exclusively used *-dup* whereas Norls used *-dup* and *-nup* at a ratio of 4 to 10, that is, Norls mostly used *-nup* (71%). Fig. 2 shows the distribution of variants in the two learning phases in different between-participant conditions.

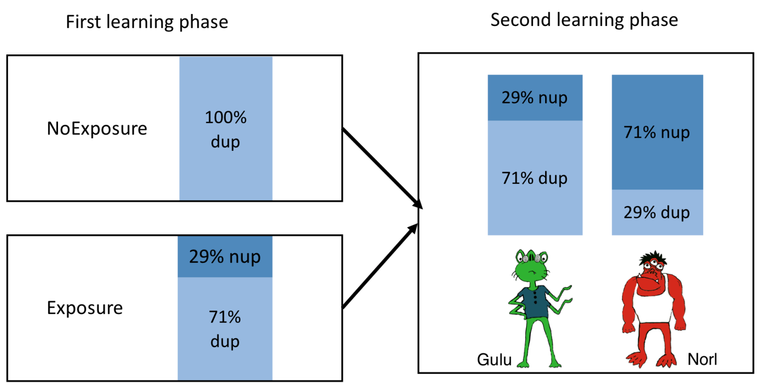

Figure 2. Distribution of variants in the two experimental conditions in Experiment 1

Results

Analyses were conducted using the R Statistical environment (R Core Team, 2014); linear models were run using the lme4 library (Bates, Mächler, Bolker, & Walker, 2015), and plots were created using ggplot (Wickham, 2016). Data and analysis scripts are available at osf.io/jeshr/.

On average, it took the participants (outliers excluded) 53 minutes ($sd = 14$) to complete the whole experiment. Out of a maximum of 168 points, participants achieved an average score of 153 ($sd = 13$).

We conducted exploratory correlation tests between score and duration, between score and participant age, and between duration and age. There was a positive correlation between duration and age (r=0.27, df=91, p=0.008), but score was not found to be correlated with any variable. A Welch two-sample t-test found a significant duration difference between Penn and Prolific subjects (t = -3.2, df = 57, p-value = 0.002), but not between subjects of different genders or conditions. No effect of gender, condition, or subject pool (i.e., Penn vs. Prolific) was found on score.

Results of suffix- and alien-selection tasks. Our primary planned analysis concerned responses in the suffix- and alien-selection tasks. Fig. 3 shows the results for the suffix-selection task (i.e., the median *-nup* selection rates by condition and alien stimulus) and Fig. 4 shows the results for the alien-selection task (i.e., the median Norl selection rates by condition and suffix stimulus). It is notable that responses did not match the frequencies in the input (where Norls used *-nup* 70% of the time). Nonetheless, the results of both tasks are consistent with our hypothesis: First, participants in the NoExposure condition (where we would expect *-nup* to cause greater surprisal) were more inclined to choose a *-nup* word for a Norl and a Norl for a *-nup* word compared with those in the Exposure condition. Second, and more importantly, the gap between the nup-selection rate given a Gulu and the nup-selection rate given a Norl was greater in the NoExposure condition than the Exposure condition, as was the gap between the Norl-selection rate given *-nup* and the Norl-selection rate given *-dup*. In other words, not only were participants in the NoExposure condition more likely to associate the right species with the right suffix, they more strongly differentiated the two species based on suffix use (and vice versa).

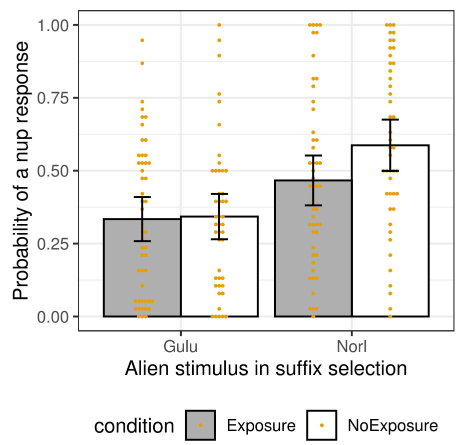

Figure 3. Mean suffix-selection task results for the two conditions of Experiment 1. Error bars indicate 95% confidence intervals. Points indicate participant means.

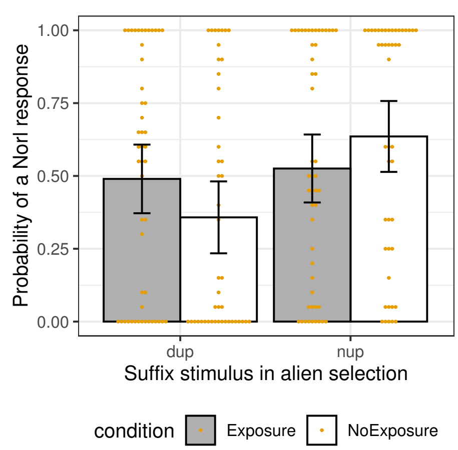

Figure 4. Mean alien-selection task results for the two conditions of Experiment 1. Error bars indicate 95% confidence intervals. Points indicate participant means.

A mixed-effects logistic regression model was conducted with Response (Suffix Selection: *-nup*; Alien Selection: Norl) as the dependent variable, Condition (Exposure-0; NoExposure-1) and Stimuli (Alien: Norl-0, Gulu-1; Suffix: *-dup*-0, *-nup*-1), as well as their interactions as independent variables, and Participant and Word as random factors. The result is shown in Table 2.

Table 2
Mixed-effects logistic regression results for Experiment 1

Tasks	**Fixed effects**	**Coef.**	**SE**	**z value**	**Pr(>|z|)**
Suffix Selection	(Intercept)	-0.20	0.13	-1.48	0.14
	Alien(Gulu)	-0.62	0.07	-8.69	<0.001
	Condition(NoExposure)	0.58	0.18	3.26	0.001
	Alien(Gulu):Condition(NoExposure)	-0.51	0.10	-4.93	<0.001
Alien Selection	(Intercept)	-0.04	0.06	-0.64	0.52
	Suffix(nup)	0.14	0.09	1.58	0.11
	Condition(NoExposure)	-0.56	0.09	-5.87	<0.001
	Suffix(nup):Condition(NoExposure)	1.05	0.13	7.81	<0.001

In suffix selection, the model revealed a significant Alien effect ($\beta = -0.62$), indicating that participants in the baseline (Exposure) condition selected *-nup* more for Norls than for Gulus, as well as a significant Condition effect ($\beta = 0.58$), indicating that participants in the NoExposure conditions selected *-nup* more overall than those in the Exposure condition. Crucially, the model also revealed a significant interaction ($\beta = -0.51$), indicating that participants in the NoExposure condition were less likely to select *-nup* for Gulus and more likely to do so for Norls than participants in the Exposure condition.

In alien selection, there was no effect of Suffix – indicating that the Norl responses were not differentiated across the two suffixes for participants in the Exposure condition – but there was a significant main effect of Condition ($\beta = -0.56$), indicating that participants in the NoExposure condition gave Norls as a response less than participants in the Exposure condition. Again, the interaction was also significant ($\beta = 1.05$), indicating that participants in the NoExposure condition were significantly more likely to select the Norl for a *-nup* word than those on the Exposure conditions.

Taken together, these results indicate that participants in the NoExposure condition built a stronger sociolinguistic association between suffixes and alien species than in the Exposure condition.

Comparison between Tasks. As discussed above, we expected that, while the results of the two tasks would point in the same direction, they would not be identical. We therefore compared them, and found inconsistencies as anticipated. The Norl selection rate in response to *-nup* was well aligned with the Gulu selection rate given *-dup* in a complementary way (Fig. 4), but the pattern for suffix selection was not as straightforward: The *-nup* selection rate given a Norl seems to be independent from the *-dup* selection rate given a Gulu (Fig. 3). To demonstrate this asymmetry between tasks, we plotted the response patterns of the two tasks for each individual participant in Fig. 5.

The left panel, showing results from the suffix-selection task, has *dup*-selection rates

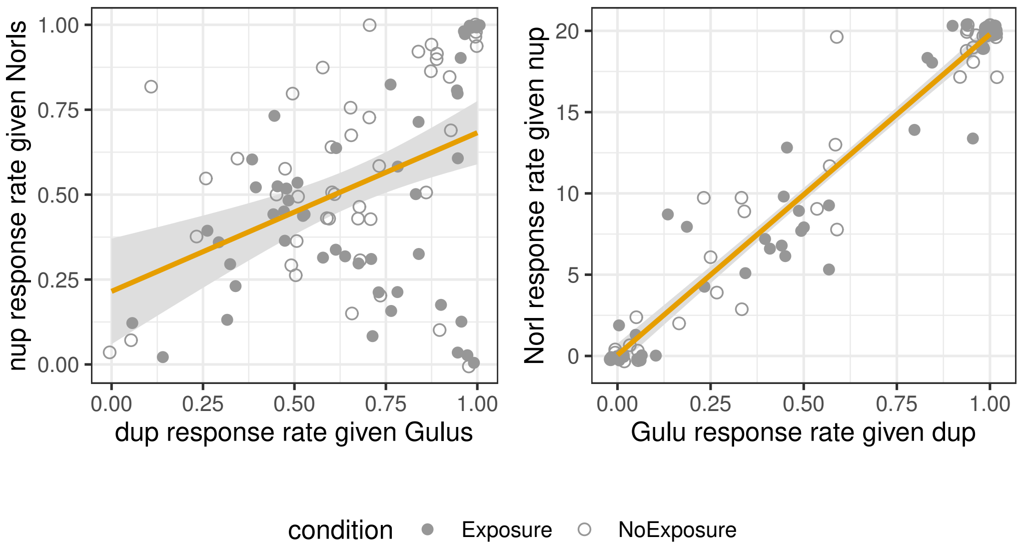

Figure 5. Task-specific response patterns of individual participants in Experiment 1. Left panel: suffix-selection task. Right panel: alien-selection task. Points indicate individual observations.

given a Gulu on the x-axis, and *nup*-selection rates given a Norl on the y-axis; the right panel, showing results from the alien-selection task, has the individual Gulu response rates given -*dup* on the x-axis, against Norl response rates given a -*nup* on the y-axis. Points indicate individual observations. Notably, in the right panel, individual observations show a strong tendency to be distributed along the diagonal of the graph, indicating that participants treated the variant suffixes as socially complementary, given this question. In addition, quite a few observations are densely distributed near the left-bottom corner and the right-top corner, indicating that the strategy established for alien selection given suffixes was rather categorical for many participants. Consistent with the above observations, a Pearson correlation test reported a very high correlation between these two response patterns ($r = 0.97, df = 91, z = 2.09, p < 0.001$)[4]

Compared with responses to the alien-selection task, responses to suffix selection did not show so strong a language-identity association, as indicated by the more scattered individual observations and a lower Pearson correlation ($r = 0.40, df = 91, z = 0.42, p < 0.001$). Based on a Fisher r-to-z transformation, difference between the two correlations was significant ($z = 11.07, p < 0.001$)

In short, it seems that participants tended in the alien-selection task to make clear, dichotomous, language–identity associations; this was considerably less apparent in the suffix-selection task.

Generalization to new words. We hypothesized that participants would apply the language-identity association they learned from the training stage to novel words they had never seen before. Fig. 6 shows participants' -*nup* responses in the suffix-selection task (left

[4]z values reported for Pearson correlation Coefficients in this paper are calculated with the Fisher r-to-z transformation: $z = 0.5 * log((1 + rho)/(1 - rho))$.

panel) and Norl responses in the alien-selection task (right panel) with seen and unseen words separated.

Figure 6. Mean suffix- and alien-selection task results for seen and unseen words in Experiment 1. Error bars indicate 95% confidence intervals. Points indicate participant means.

At a quick glance, responses to seen and unseen stimuli strongly resembled each other in both conditions and both tasks. Responses to unseen stimuli also replicated the difference in sociolinguistic acquisition between conditions, in that the responses in the NoExposure condition were better aligned with the expected associations (Gulu–*dup*, Norl–*nup*) in both tasks, compared with those in the Exposure condition.

A mixed-effects model was conducted for each of the two tasks, with Response (Suffix or Alien) as the dependent variable, Participant and Word as random factors, and Condition, Stimulus (either Alien or Suffix), Novelty and their interactions as fixed effects. The results showed no significant Novelty effect ($\beta = 0.23$ in suffix selection; $\beta = 0.10$ in alien selection, p >0.05), nor were there any Novelty-relevant two-way interactions. There was, however, a significant three-way interaction for suffix selection ($\beta = -1.03, p < 0.001$): Participants in the NoExposure condition were more likely to choose *-nup* in response to a Norl and less likely to choose it in response to a Gulu on unseen words than seen words (see the left panel of Fig. 6). There was no such interaction for the alien-selection task (see Supplementary Material: Tables 1–2). These results indicate that the acquired language-alien associations can be generalized to new lexical items, and the first-encounter benefit in sociolinguistic learning applies to both familiar and unfamiliar words.

Discussion of Experiment 1

We predicted that participants would better learn to associate a linguistic variant with the social group that most used it if the variant violated expectations by being entirely novel. This prediction was supported. That is, participants in the NoExposure condition not only associated *-nup* with Norls and Norls with *-nup* more than participants in the Exposure condition, but they were also less likely to associate *-dup* with Norla and Gulus with *-nup*. This suggests that the first encounter with a novel variant facilitated sociolinguistic acquisition of that variant. That being said, it should be noted that the general *-nup* response rate for Norls was relatively low, compared with the ratio of Norl–*nup* co-occurence in the input (71%).

It is also notable that the suffix-selection results and alien-selection results exhibited different patterns. In alien selection, the Norl selection rate for *nup*-words was well aligned with the Gulu selection rate for *dup*-words, as indicated by the complementary ratios of Norl responses for *dup*-words and for *nup*-words in both groups of participants (Fig. 3). However, this symmetric response pattern did not hold in suffix selection: The ratio of -*nup* chosen in response to a Norl did not necessarily predict the ratio of -*dup* chosen in response to a Gulu (Fig. 4). One way of interpreting the former pattern is to say that, in selecting who might have used a particular form, participants tended to adopt a dichotomous language-identity ideology. That is, they treated -*nup* as a distinctively Norl variant and *dup* as a distinctively Gulu variant in the same way that an English person asked which group of speakers they associate with a velar fricative pronunciation of /k/ would likely mention Liverpool. Such an ideology was apparently not adopted in suffix selection. One potential interpretation of this would be that different processing mechanisms are employed for the two tasks, but there are simpler interpretations. In particular, the asymmetry in task responses may reflect an asymmetry in the input data. Here it is helpful to return to the natural-language analogy: The pronunciation [bʊx] for *book* is rare outside the Merseyside area, so is very strongly associated with its inhabitants. It is not the case, however, that all people from Merseyside pronounce the word in this manner. Similarly, in the input data in our experiment, only Norls were ever seen using the -*nup* suffix, so it was relatively straightforward to infer that any example of that suffix would have come from a Norl. On the other hand, while no Gulus ever used -*nup*, Norls did sometimes use -*dup*, so it was less straightforward to infer, given a Norl, which suffix it might have used. This does not fully account for the almost categorical selection of Gulu given the -*dup* suffix in the alien-selection task, or the number of participants who thought a Gulu might have used -*nup* in the suffix-selection task, but it likely plays a role in explaining the asymmetry nonetheless.

It is worth discussing the extension of associations to unseen words in the same light. On the whole the pattern here was the same as for seen words. However, for the suffix-selection task in the NoExposure condition only, there was a stronger tendency for participants to treat -*nup* as a distinctly Norl variant. Participants had seen Norls using both -*nup* and -*dup* a number of times, so there was evidence that for seen words, Norls varied. However, for *unseen* words there was no direct evidence that Norls would vary, so participants' inference about sociolinguistic associations between variant and species had more of a role to play. It is interesting, based on this interpretation, that the same thing did not occur for the alien-selection task. However, as discussed, the input data were, with respect to this task, much more univocal: the suffix -*nup* was only ever used by Norls.

Finally, Experiment 1 investigated a rather dramatic violation of expectation – exposure to an entirely new variant should be expected to be rather salient to learners. It would be interesting, therefore, to investigate whether the effect we found in Experiment 1 extends to violations of expectation concerning an already established variant. This is the focus of Experiment 2.

Experiment 2: Constraint Violation

Experiment 2 used a similar paradigm to Experiment 1, except that we modified the suffixation patterns to investigate a different source of surprisal. Instead of surprisal caused by encountering a variant for the first time (as in Experiment 1), Experiment 2 investigated whether surprisal caused by encountering a linguistic variant in an apparently ungrammatical context would also facilitate sociolinguistic acquisition.

Method

Participants. 103 participants completed Experiment 2 online within 1.5 hours. After we excluded participants whose duration was below the 2.5% quantile or above the 97.5% quantile of all participants, there were 100 participants left whose data were used for the final analysis. They were 71 females and 29 males, aged 17–78 (mean: 29.3) years old. 30 of them were recruited from the subject pool of the University of Pennsylvania, and the remaining 70 were recruited through the Prolific Academic website. There were 50 participants in the *Conditioned* condition and 50 in the *Unconditioned* condition.

Materials and Procedure. The same words and aliens were used in Experiment 2 as in Experiment 1. The procedure was also the same, consisting of a two-stage learning phase and a test phase with two tasks.

Between-participant Conditions. There were two between-participant conditions based on the linguistic environment for the suffix *-nup*. In the *Conditioned* condition participants in the first learning stage only ever saw *-nup* attached to the four nasal-ending stems (i.e., *falon*, *hiwen*, *tugan* and *wukin*), whereas *-dup* was attached to the 10 stems that did not end in a nasal. By contrast, participants in the *Unconditioned* condition were exposed to the two suffix variants in free variation (i.e., both *-nup* and *-dup* occurred with both nasal and non-nasal stems), though the variants still occurred at a ratio of ten (*-dup*) to four (*-nup*) just as in the Conditioned condition. In the second learning phase, Gulus exhibited precisely the suffixation pattern of the first stage (in terms of both ratio and conditioning), whereas Norls used the two suffixes freely across contexts at a ratio of four (*-dup*) to ten (*-nup*).

Fig. 7 shows the distribution of variants in the two learning phases in different between-participant conditions. It should be noted that participants in both conditions would see Gulus and Norls using *-dup* and *-nup* with a reversed ratio: 10 to 4 and 4 to 10. The crucial difference lay in the fact that, in the *Conditioned* condition, participants would experience apparently ungrammatical words from *Norls* but participants in the *Unconditioned* condition would not.

Similar to Experiment 1, we predicted that participants in the Conditioned condition would experience greater surprisal when they saw Norls using the two variants, especially *-nup*, in an ungrammatical way, and would be facilitated by this surprisal in learning the association between *-nup* and the Norl species, compared with participants in the Unconditioned condition.

Results

On average participants took 51 minutes to complete Experiment 2 (sd = 13) and achieved 154 (sd = 12) points out of 168.

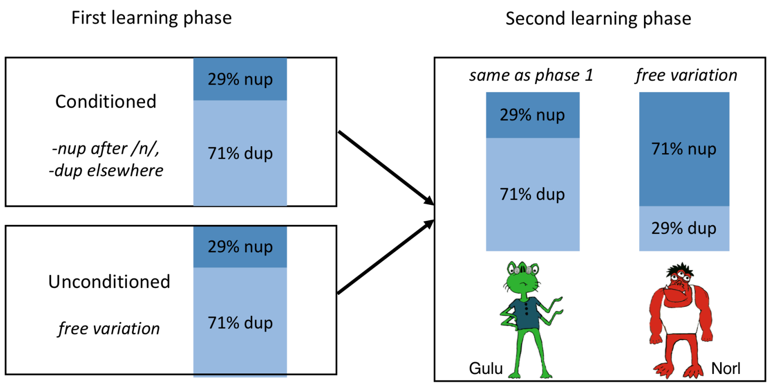

Figure 7. Distribution of variants in the two between-participant conditions in Experiment 2

We conducted exploratory correlation tests between score and duration, between score and participant age, and between duration and age. Again, a positive correlation was found between duration and age (r=0.31, df=97, p=0.002), but score was not found to be correlated with any variable. A Welch two-sample t-test found no effect of gender, condition, or subject pool on score or duration.

Results of suffix-selection and alien-selection tasks. Fig. 8 shows the aggregate results of suffix-selection and alien-selection tasks in Experiment 2.

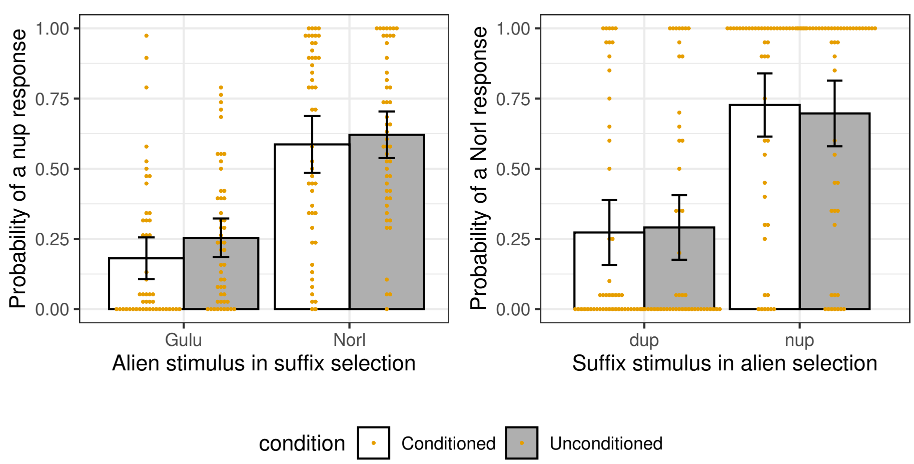

Figure 8. Mean results for suffix-selection task (left) and alien-selection task (right) in Experiment 2. Error bars indicate 95% confidence intervals. Points indicate participant means.

Both panels show a clear stimulus effect: Participants in both groups and both tasks tended to associate Norls more with *-nup* and Gulus more with *-dup*. Regarding the predicted difference between conditions, the results were more mixed. On the one hand, the probability of selecting a Norl in response to each of the two suffixes was rather similar across the two conditions. On the other hand, the difference between the “incorrect” association rate (i.e., associating Gulus with *-nup* and Norls with *-dup*) and the "correct" association rate (i.e., associating Norls with *-nup* and Gulus with *-dup*) was greater in the Conditioned condition than in the Unconditioned condition (0.41 vs. 0.36 for alien selection and 0.48 vs. 0.41 for suffix selection), in line with our predictions.

We conducted mixed-effects logistic models with Response (Suffix selection: *-nup*; Alien selection: Norl) as the dependent variable, Participant and Word as random intercepts, and Condition (Conditioned-0, Unconditioned-1), Stimulus (Alien: Norl-0, Gulu-1; Suffix: *-dup*-0, *-nup*-1) and their Interaction as independent variables. The result, as shown in Table 3, revealed a similar result across tasks, showing a significant effect for Stimulus in the baseline condition (Alien: $\beta = -1.67, p < 0.001$; Suffix: $\beta = 1.71, p < 0.001$) and a significant interaction between Stimuli and Condition (Alien:Condition: $\beta = -0.41, p < 0.001$; Suffix:Condition: $\beta = 0.33, p = 0.02$), but no main effect of Condition alone.

Table 3
Mixed-effects logistic regression results for Experiment 2

Tasks	**Fixed effects**	**Estimate**	**Std.Error**	**z value**	**Pr(>|z|)**
Suffix selection	(Intercept)	0.50	0.13	3.76	<0.001
	Alien(Gulu)	-1.67	0.08	-22.27	<0.001
	Condition(Conditioned)	-0.15	0.18	-0.86	0.39
	Alien(Gulu):Condition(Conditioned)	-0.41	0.11	3.74	<0.001
Alien selection	(Intercept)	-0.88	0.07	-12.62	<0.001
	Suffix(nup)	1.71	0.10	17.49	<0.001
	Condition(Conditioned)	-0.14	0.10	1.44	0.15
	Suffix(nup):Condition(Conditioned)	0.33	0.14	2.35	0.02

Again, the significant Stimulus effect indicates that participants selected *-nup* more for Norls than for Gulus and selected Norls more for *-nup* than for *-dup*. Crucially, the significant interaction suggests that the differences in Response associated with Stimulus were significantly larger in the Conditioned condition than in the Unconditioned condition in both tasks, indicating that participants in the Conditioned condition established stronger associations between aliens and suffixes.

Comparison between tasks. As in Experiment 1, we found an asymmetry between the responses in different tasks, as shown in Fig. 9.

The left panel shows the relationship between the *-dup* selection rate given a Gulu and the *-nup* rate given a Norl, and the right panel shows the relationship between the Gulu selection rate given a *-dup* and the Norl rate given *-nup*. The results look very similar to those in Experiment 1, indicating that participants' alien-selection patterns tended to be categorical, and that a participant's Norl selection rate given *-nup* is almost always predictable from their Gulu selection rate given *-dup*. Accordingly, the Pearson correlation between these dimensions is very high ($r = 0.98, df = 98, z = 2.30, p < 0.001$). In comparison, observations in the left panel are distributed much less closely to the diagonal; participants tended to estimate the probability of different variances independently for the two alien species. The Pearson correlation was also lower ($r = 0.52, df = 98, z = 0.58, p < 0.001$). Again, the difference between correlations was significant ($z = 11.86, p < 0.001$).

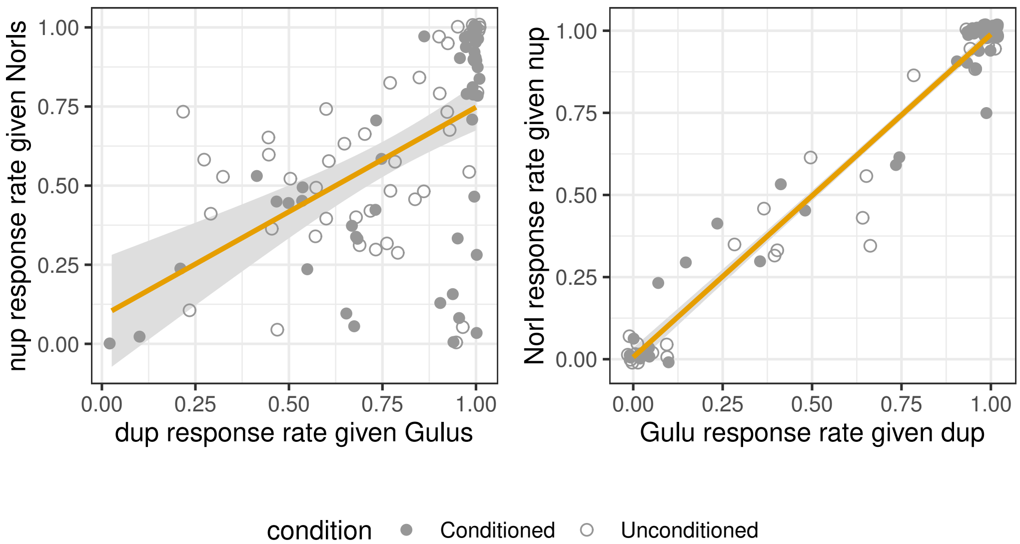

Figure 9. Task-specific response patterns of individual participants in Experiment 2. Left panel: suffix-selection task. Right panel: alien-selection task. Points indicate individual observations.

Learning performance. The interaction that we found between stimulus and condition is consistent with our hypothesis. However, the results in Experiment 2 were not as convincing as in Experiment 1: The effect sizes were smaller and there was no main effect of condition. A possibility worth considering in evaluating these results is the role played by differences in learning performance. It seems a priori clear that we should expect our predicted effect only if participants learned the grammatical rule that was apparently broken by Norls in the second training phase. If they did not, there was no expectation to be violated. To investigate this possibility, we conducted a post hoc analysis in which we took the scores earned in the training phase as a proxy for learning performance and divided participants into “good learners” and “poor learners” in each condition according to whether their score was above or below the group median. We then asked whether results differed for the two groups. Fig. 10 shows the suffix-selection and alien-selection results for the 50 “good” and 50 “poor” learners in the different conditions.

Several differences can be seen between the two groups of participants in their responses. First, good learners showed a higher *-nup* rate for Norls and a lower *-nup* rate for Gulus in suffix selection, as well as a higher “Norl” rate for *-nup* and a lower one for *-dup*, compared with poor learners. In other words, the responses of good learners were better aligned with the sociolinguistic association in the input than those of poor learners. Second, the pattern of results for good learners, but not poor learners, is consistent with the predicted learning facilitation effect: In suffix selection, although participants in the two conditions showed comparable *-nup* rates in response to Norls, participants in the Conditioned condition exhibited a lower *-nup* rate for Gulus; in alien selection, participants in the Conditioned condition had a higher Norl selection rate for *-nup* words and a lower Norl rate for *-dup* words compared with those in the Unconditioned condition. This expected between-condition dif-

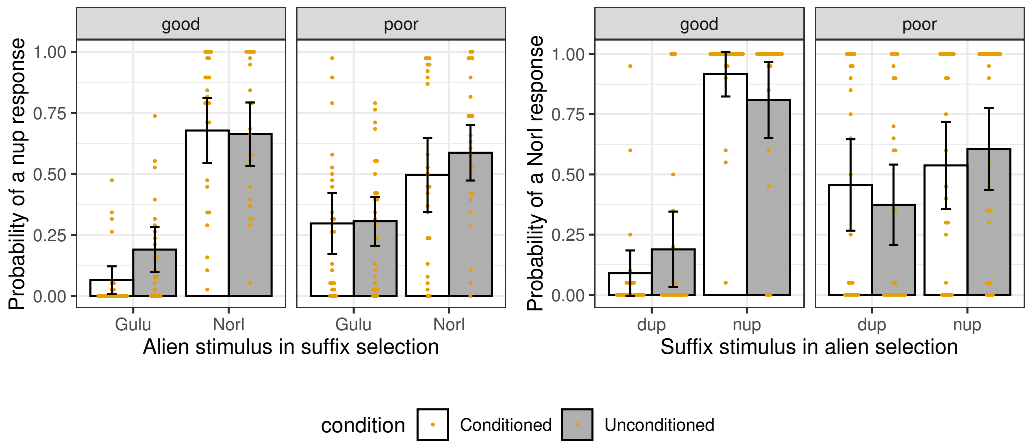

Figure 10. Mean results for suffix- and alien-selection tasks by "good" and "poor" learners in Experiment 2. Error bars indicate 95% confidence intervals. Points indicate participant means.

ference was in fact reversed in the results for poor learners. We replicated mixed-effects logistic regression on the data of good learners. The results show significant effects of Alien ($\beta = 3.65$), Condition ($\beta = 1.42$) and their interaction ($\beta = -1.52$) in suffix selection, as well as significant effects of Suffix ($\beta = 4.01$), Condition ($\beta = 0.62$) and Interaction ($\beta = -1.21$) in the alien-selection task (See Supplementary material, Tables 3–4).

For comparison, we performed the same analysis on the data in Experiment 1. Again, the results show that "good" and "poor" learners differed in how well their responses were aligned with the alien-suffix association in the input. Unlike experiment 2, however, the between-condition difference was maintained in both learner groups. See Supplementary materials, Fig. 1, for the results of different learner groups in Experiment 1.

A more direct way to evaluate the influence of rule learning is to examine whether participants' sociolinguistic acquisition in the Conditioned condition was correlated not with their overall learning success, but with the degree to which they used the rule correctly. For each participant, we measured this by counting how many of their answers conformed to the rule (i.e., how often they paired *-dup* with a non-nasal-final and *-nup* with an *n*-final word given a Gulu) and how many of their answers violated the rule (i.e., how often they paired *-dup* with an *n*-final word or *-nup* with a non-nasal-final word given a Gulu, and so on), subtracting the latter from the former. Based on this, rule adherence was indeed positively correlated with the *-nup* response rate given a Norl ($r = 0.57, df = 48, z = 0.65, p < 0.001$) and (more strongly) with the Norl response rate given *-nup* ($r = 0.78, z = 1.05, df = 48, p < 0.001$).

Generalization to new words. Since the aggregate pattern of results for Experiment 2 is different from both the pattern for good learners and the pattern for bad learners, we consider "good" and "bad" learners separately when evaluating the generalization of acquisition to new words. Fig. 11 shows the results of suffix selection (left) and those of alien selection (right) by the learning performance of participants (vertically) and by the novelty of stimuli (horizontally).

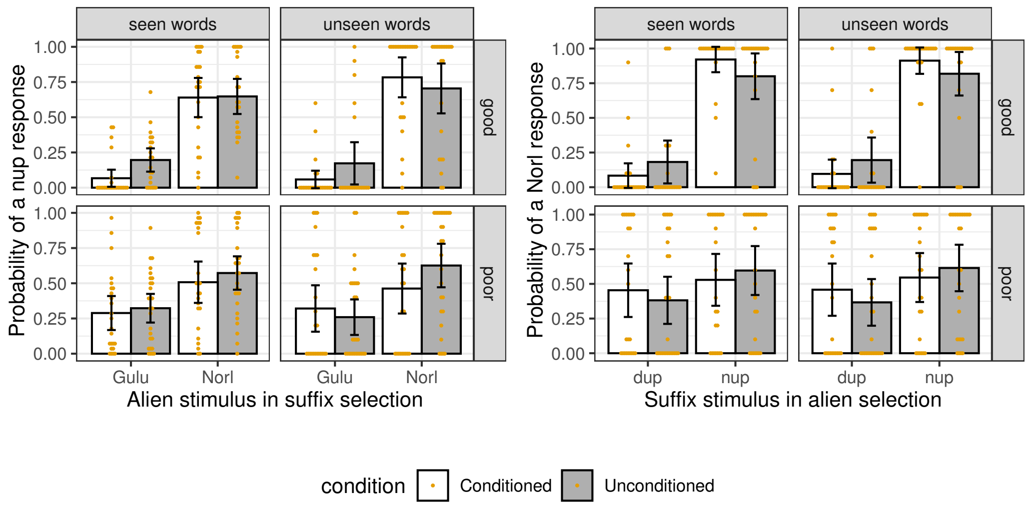

Figure 11. Mean results of suffix- and alien-selection tasks for seen and unseen words in Experiment 2 by "good" and "poor" learners. Error bars indicate 95% intervals. Points indicate participant means.

Although learners with different performance showed distinct patterns from each other, the behaviors with seen and unseen stimuli were highly consistent within each of the two learner groups. "Good learners" showed the correct alien-language association as well as additional facilitation from rule violation with both seen and unseen stimuli. "Poor learners" also showed consistent behaviors across seen and unseen stimuli, although the choices they made were mostly around the chance level.

Two mixed-effects models were conducted, one on each task, with Response (Suffix or Alien) as the dependent variable, Participant and Word as random factors, and Condition, Stimuli (Suffix or Alien), Novelty and their interactions as mixed effects. The results for the suffix-selection task showed significant interactions for Condition:Novelty ($\beta = -0.45$) and Alien:Condition:Novelty ($\beta = 0.49$) but the results for the alien-selection task showed no significant Novelty effect or Novelty-relevant interactions. (Supplementary material: Tables 5–6). This is consistent with the pattern we observed in Experiment 1 (Fig. 6).

Discussion of Experiment 2

We predicted that acquisition of sociolinguistic meaning would be facilitated for participants in the Conditioned condition, compared with those in the Unconditioned condition, because of expectation violation caused by observing a variant in an apparently ungrammatical context. We found a significant interaction in line with this prediction, but for the results as a whole the support for our hypothesis was not as strong as in Experiment 1. A post hoc analysis revealed differences in the pattern of results depending on how well participants learned the language, with above-average learners in our sample showing the predicted effect, consistent with the view that grammatical rule violation should facilitate acquisition of the sociolinguistic pattern only among individuals who have learned the grammatical rule. This was a post hoc analysis so should be treated with caution; however, its results are suggestive and warrant attention in future work.

The results of Experiment 2 also replicate those of Experiment 1 in showing different response patterns associated with the two tasks (selecting an alien given a suffix vs. selecting a suffix given an alien), as well as an ability to generalize patterns, whether accurate or inaccurate, to unseen words.

General Discussion

We hypothesized that violation of expectation would cause a linguistic variant to be more salient and that, as a result of this, an association between this variant and a particular social group would be easier to learn. We tested this hypothesis in two experiments, each investigating a different kind of expectation violation. The first experiment investigated exposure to a previously unencountered variant while the second investigated exposure to a variant that had previously occurred within a narrower grammatical context. In the first experiment the expectation violation had the predicted effect: Participants were more likely to associate the new suffix with the correct alien species (and the correct alien species with the new suffix) when the suffix had not been encountered in the initial training phase. We also found that this effect extended to previously unseen words. In the second experiment, we found evidence for the predicted effect, but not as convincingly as in Experiment 1. However, when we divided participants into "good learners" and "poor learners" (whether these are defined generally in terms of performance on the learning task or more specifically in terms of performance at reproducing the grammatical rule), we found a stronger effect for the former. The division into good and poor learners has a clear precedent in earlier work (Rácz et al., 2017) and makes good theoretical sense: That is, we should not expect violation of a grammatical rule to be salient to participants who have not learned that rule. (Indeed, it would be problematic for our hypothesis if we had found an effect for learners who had not acquired the rule.) All the same, we must be cautious here. The subsetting by learning performance was not planned and there may be confounds. For instance, it might well be that participants who were better at learning the constraint also happened to be better at learning the social–linguistic association. An obvious focus for future work should be to try to replicate this effect, while controlling for such issues.

An important unrelated point to be raised concerns asymmetry in the sociolinguistic patterns that participants acquired. We saw this in two related ways. First, the results looked different depending on whether we asked participants to select a suffix based on an alien stimulus or asked them to select an alien based on a suffix stimulus. Second, the pattern of association of the *-dup* suffix with Gulus was not the complement of the pattern of association of the *-nup* suffix with Norls (though, for alien-selection, the two patterns were rather complementary). These patterns are consistent with natural-language behavior. The fact that a particular group prefers a particular form is not the same as saying that a particular form is used primarily by that group. This is closely related to the point raised in Footnote 1 on page 1: The fact that pronouncing the English *-ing* suffix as *-in* has particular social meanings does not mean that pronouncing it as *-ing* has complementary meanings (Campbell-Kibler, 2007). There was also an asymmetry built into the input data – a necessary result of the social–linguistic associations in the input being non-categorical. That is, the Norls used both suffixes in all conditions, while the Gulus used both suffixes in both conditions of the second experiment. The alien-selection task therefore involved a different kind of inference from the suffix-selection task. This is also consistent with natural-language patterns, and the use of gradient rather than categorical associations in the input data serves the goal of ecological validity, while also shedding light on the threshold at which social–linguistic associations can be acquired.

It is interesting to speculate on what "learning" means in the context of our experiments: Are you a better learner if you cement an association between an alien and a suffix or if you recognise and reproduce weighted variation? In the NoExposure condition of Experiment 1, participants responded to the Gulu/Norl asymmetry by essentially frequency-matching the Norls' 30%-70% pattern. They did not deterministically associate -dup with the Gulus, even though this would have been consistent with the input. This might be because participants in a forced-choice task will always exploit all available choices, more so than in an open response task (see, e.g., Treiman, Seidenberg, and Kessler 2015). It might also be because people are generally poor at taking background probabilities into account (see, e.g., Cruz, Baratgin, Oaksford, and Over 2015), so that the input variation for the *suffix* might be generalised for both alien species (even if one alien shows 100%–0% in training). The "good" learners for Experiment 1 were consistent with the overall pattern. For Experiment 2, however, the good learners overshot the training pattern and were closer to categorical, especially for the alien-selection task. One possible explanation is that the task in Experiment 2 was simply quite a lot harder, resulting in regularizing behavior (cf. Hudson Kam & Chang, 2009). If this is right, a more nuanced account of sociolinguistic acquisition, based on our data, would state that expectation violation can facilitate learning, but that what learning means depends on the nature and complexity of the data.

A related question concerns whether sociolinguistic acquisition is simply a special case of broader patterns of associative learning, at least insofar as expectation violation is involved (see, e.g., Hirshman, 1988; Nixon, 2020; Thorwart & Livesey, 2016). Our study did not compare different domains of learning, but we would speculate that analogous results would be likely in a non-linguistic task. That being said, it might well be that what counts as a violation of expectation varies between domains. On a similar note, for instance, the difference between spatial orientation and gender observed by Rácz et al. (2017) might disappear or be reversed in other domains. We do not, however, consider that there is any strong reason to assume that sociolinguistic acquisition as investigated in this study is fundamentally different from other forms of associative learning.

There are inevitable limitations of our approach. One concerns the written modality we employed. This made our stimuli easier to create and our data easier to record and analyze. It also avoided potential nuisance (or even extraneous) variables brought about by misattribution of variation to acoustic effects. However, in real life, the kind of variation we investigated would of course be more likely to occur in speech. Future work should take this into account and investigate a wider range of modalities. The kind of task in which our participants engaged also differed in important ways from the acquisition of social–linguistic associations in the real world. The experiment did not, for instance, involve communication; the alien species, moreover, likely did not have much practical meaning for participants, in the sense that real-life social distinctions such as socioeconomic class, geographical origins, or race often have very practical consequences for day-to-day life. (For a paradigm that allows both issues to be somewhat mitigated, see Sneller and Roberts 2018.) Nor is it usual to acquire linguistic experience divorced from sociolinguistic context, as in our first learning phase. These compromises with respect to ecological validity afforded us greater experimental control, and we are not concerned that they seriously undermine our results. It is nonetheless possible that a richer and more practically meaningful learning context might have increased participant motivation and reduced the number of poor learners (on the role of motivation in adult language acquisition generally, see Dörnyei and Schmidt 2001). This would be an independently interesting question to investigate further. It is also worth noting, however, that while our learning paradigm has many differences from more naturalistic acquisition, it bears several resemblances to popular language-learning apps. An interesting further avenue to pursue might be to investigate whether our results could have applications in the teaching of sociolinguistic variation in second-language learning. Another obvious avenue to pursue would be to investigate in a more quantifiable way the role of different degrees of violation and surprisal. Our work looked at two different kinds of expectation violation that differed intuitively in terms of the degree of likely surprisal. That is, the complete absence of a variant in prior experience feels like a stronger contrast with current experience than the absence of that variant *only in certain grammatical contexts*. Future work, following Jaeger and Weatherholtz (2016), could quantify this and more systematically investigate the consequences of different degrees of expectation violation.

References

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, *39*(4), 437–456. doi: https://doi.org/10.1017/s0047404510000400

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: https://doi.org/10.1007/0-387-22747-4_4

Bell, A. (2001). Back in style: Reworking audience design. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 139–69). Cambridge: Cambridge University Press. doi: https://doi.org/10.1017/cbo9780511613258.010

Beňuš, Š. (2014). Social aspects of entrainment in spoken interaction. *Cognitive Computation*, *6*(4), 802–813. doi: https://doi.org/10.1007/s12559-014-9261-4

Campbell-Kibler, K. (2007). Accent, (ING), and the social logic of listener perceptions. *American speech*, *82*(1), 32–64. doi: https://doi.org/10.1215/00031283-2007-002

Campbell-Kibler, K. (2010). The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, *22*(3), 423–441.

Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, *32*(1), 111–140. doi: https://doi.org/10.1016/s0095-4470(03)00009-3

Cruz, N., Baratgin, J., Oaksford, M., & Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in psychology*, *6*, 192.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329. doi: https://doi.org/10.1016/j.cognition.2011.10.017

Docherty, G. J., & Foulkes, P. (2005). Glottal variants of /t/ in the Tyneside variety of English. In W. J. Hardcastle & J. M. Beck (Eds.), *A figure of speech: A festschrift for John Laver* (pp. 173–197). London: Lawrence Erlbaum. doi: https://doi.org/10.4324/9781410611888

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

Dörnyei, Z., & Schmidt, R. (2001). *Motivation and second language acquisition* (Vol. 23). Honolulu: Second Language Teaching and Curriculum Center.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, *34*(4), 409–438. doi: http://dx.doi.org/10.1016/j.wocn.2005.08.002

Foulkes, P., Docherty, G. J., Khattab, G., & Yaeger-Dror, M. (2010). Sound judgements: Perception of indexical features in children's speech. In D. Preston & N. Niedzelski (Eds.), *A reader in sociophonetics* (pp. 327–356). Berlin: Walter De Gruyter. doi: https://doi.org/10.1515/9781934078068.3.327

Foulkes, P., Docherty, G. J., & Watt, D. (2005). Phonological variation in child-directed speech. *Language*, *81*(1), 177–206.

Fraundorf, S. H., & Jaeger, T. F. (2016). Readers generalize adaptation to newly-encountered dialectal structures to other unfamiliar structures. *Journal of Memory and Language*, *91*, 28–58. doi: https://doi.org/10.1016/j.jml.2016.05.006

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the association for computational linguistics on language technologies* (pp. 1–8). doi: https://doi.org/10.3115/1073336.1073357

Harrington, J., Kleber, F., Reubold, U., Schiel, F., & Stevens, M. (2018). Linking cognitive and social aspects of sound change using agent-based modeling. *Topics in Cognitive Science*, *10*(4), 707–728. doi: https://doi.org/10.1111/tops.12329

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484. doi: https://doi.org/10.1016/

j.wocn.2005.10.001

Hespos, S. J., & Baillargeon, R. (2008). Young infants' actions reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings. *Cognition*, *107*(1), 304–316. doi: https://doi.org/10.1016/j.cognition.2007.07.009

Hirshman, E. (1988). The expectation-violation effect: Paradoxical effects of semantic relatedness. *Journal of Memory and Language*, *27*(1), 40–58. doi: https://doi.org/10.1016/0749-596X(88)90047-2

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 815.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66. doi: https://doi.org/10.1016/j.cogpsych.2009.01.001

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10-12), 1489–1506.

Jaeger, T. F., & Weatherholtz, K. (2016). What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology*, *7*. doi: https://doi.org/10.3389/fpsyg.2016.01115

Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General*, *133*(3), 450. doi: http://dx.doi.org/10.1037/0096-3445.133.3.450

Kerswill, P., & Williams, A. (2002). "salience" as an explanatory factor in language change: Evidence from dialect levelling in urban England. *Contributions to the Sociology of Language*, *86*, 81–110. doi: https://doi.org/10.1515/9783110892598.81

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1989). The child as linguistic historian. *Language Variation and Change*, *1*(1), 85–97. doi: https://doi.org/10.1017/s0954394500000120

Labov, W. (2001). *Principles of Linguistic Change. Volume 2: Social Factors.* Malden, MA: Blackwell.

Labov, W., Ash, S., Ravindranath, M., Weldon, T., Baranowski, M., & Nagy, N. (2011). Properties of the sociolinguistic monitor 1. *Journal of Sociolinguistics*, *15*(4), 431–463. doi: https://doi.org/10.1111/j.1467-9841.2011.00504.x

Lambert, W. E. (1967). A social psychology of bilingualism. *Journal of Social Issues*, *23*(2), 91–109. doi: https://doi.org/10.1111/j.1540-4560.1967.tb00578.x

Levon, E. (2006). Hearing "gay": Prosody, interpretation, and the affective judgments of men's speech. *American Speech*, *81*(1), 56–78. doi: https://doi.org/10.1215/00031283-2006-003

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi: https://doi.org/10.31234/osf.io/byhxq

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751. doi: https://doi.org/10.1016/s0042-6989(03)00237-2

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312. doi: https://doi.org/10.1006/jmla.1997.2543

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, *34*(2), 202–240. doi: https://doi.org/10.1016/j.wocn.2005.05.003

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of language and Social Psychology*, *18*(1), 62–85. doi: https://doi.org/10.1177/0261927x99018001005

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081. doi: https://doi.org/

10.1016/j.cognition.2019.104081

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, *18*(1), 10–30. doi: https://doi.org/10.1177/0261927x99018001002

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology*, *8*. doi: https://doi.org/10.3389/fpsyg.2017.00051

Roberts, G. (2010). An experimental study of social selection and frequency of interaction in linguistic diversity. *Interaction Studies*, *11*(1), 138–159. doi: https://doi.org/10.1075/bct.45.08rob

Roberts, G. (2013). Perspectives on language as a source of social markers. *Language and Linguistics Compass*, *7*(12), 619–632. doi: https://doi.org/10.1111/lnc3.12052

Roberts, J., & Labov, W. (1995). Learning to talk Philadelphian: Acquisition of short a by preschool children. *Language Variation and Change*, *7*(1), 101–112. doi: https://doi.org/10.1017/s0954394500000910

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. doi: https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Smith, J., Durham, M., & Fortune, L. (2009). Universal and dialect-specific pathways of acquisition: Caregivers, children, and t/d deletion. *Language Variation and Change*, *21*(1), 69–95. doi: https://doi.org/10.1017/s0954394509000039

Sneller, B., & Roberts, G. (2018). Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition*, *170C*, 298–311. doi: https://doi.org/10.1016/j.cognition.2017.10.014

Squires, L. (2014). Social differences in the processing of grammatical variation. *University of Pennsylvania Working Papers in Linguistics*, *20*(2), 20.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94. doi: https://doi.org/10.1126/science.aaa3799

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of language and social psychology*, *18*(1), 86–100. doi: https://doi.org/10.1177/0261927x99018001006

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*, 1015.

Thorwart, A., & Livesey, E. J. (2016). Three ways that non-associative knowledge may affect associative learning processes. *Frontiers in Psychology*, *7*, 2024. doi: https://doi.org/10.3389/fpsyg.2016.02024

Treiman, R., Seidenberg, M. S., & Kessler, B. (2015). Influences on spelling: evidence from homophones. *Language, Cognition and Neuroscience*, *30*(5), 544–554.

Trudgill, P. (2008). Colonial dialect contact in the history of european languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, *37*(2), 241–254. doi: https://doi.org/10.1017/s0047404508080287

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* New York: Springer-Verlag. Retrieved from http://ggplot2.org

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. doi: https://doi.org/10.1016/s0042-6989(99)00163-7