LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Not All Indexical Cues Are Equal: Differential Sensitivity to Dimensions of Indexical Meaning in an Artificial Language

Péter Rácz,[a,b] Jennifer B. Hay,[b] and Janet B. Pierrehumbert[b,c]

[a]Central European University, [b]University of Canterbury, and [c]University of Oxford

Abstract: In this study, we investigated the learning of indexical features by English-speaking adults using a novel experimental paradigm. In a conceptual replication of Rácz, Hay, and Pierrehumbert (2017), participants learned an allomorphy pattern cued by a given social context. The social contexts were represented by conversation partners who differed by age, ethnicity, and/or gender and were positioned in various ways. The results showed that, after training, the participants were able to learn that different types of conversation partners prefer different types of allomorphs but that learning and generalization hinged on the social relevance of the cue represented by the conversation partner. These results suggest that the relevance of cues in an individual's past social

The authors would like to thank Chun Liang Chan, Patrick LaShell, Clay Beckner, Vica Papp, Jeremy Needle, Robert Fromont, and Seán Roberts. All faults remain ours. The roof runner experimental platform was developed by Chun-Liang Chan and the visuals created by Kayo Takasugi. This work was supported by a grant from the John Templeton Foundation (Award ID 36617) to JBP and JH, and a Royal Society of New Zealand Rutherford Discovery Fellowship (Grant No. E5909) to JH. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Correspondence concerning this article should be addressed to Péter Rácz, Cognitive Development Center, Department of Cognitive Science, Central European University, Október 6. u. 7, 1051 Budapest, Hungary.
E-mail: raczp@ceu.edu

DOI: 10.1111/lang.12402

experience influences their storage and learnability even at very early stages of learning a word pattern.

Keywords sociolinguistics; exemplar theory; morphology; gamification; learning

Introduction

Many words carry statistical associations with nonlinguistic contexts—they might be used more by some types of speakers than by others, or be preferred with certain interlocutors, or in certain contexts. In New Zealand English, for example, a female speaker is statistically more likely to produce the word *lovely* than is a male speaker, and a speaker who is young is unlikely to produce the word *confectionery* (Hay, Walker, Sanchez, & Thompson, 2019). Such indexical associations (Silverstein, 2009) can be remarkably complex and long-lasting and are learned and produced by language users with ease (for an overview, see Hay, 2018). Nonrandom associations between linguistic and nonlinguistic contexts can play a crucial role in early word learning (Woodward & Markman, 1998) and continue to influence language processing throughout the lifespan (Chater & Manning, 2006).

Despite their relevance to language use, language learning, and language processing, the development of these associations is poorly understood. Existing results come from two areas. On the one hand, a body of experimental work exists on the role of the context in category learning in general, especially in visual processing (see, e.g., Borji & Itti, 2013). On the other hand, researchers have discovered a lot about the ways in which context is indexed in linguistic conventions and how this contributes to language variation and change (see, e.g., Bucholtz, 1999; Gudmestad, 2012; Hay & Drager, 2010; Niedzielski, 1999).

The missing piece is how people learn these associations between language and the nonlinguistic context—the subject of this article. We used an artificial language learning task to investigate how indexical associations in language are learned. The great benefit of this paradigm is that it allowed us to operationalize an otherwise rich and complex problem. Indeed, artificial language tasks have been used to great effect in studying language, its evolution, and its variation (Kirby, Griffiths, & Smith, 2014; Roberts, 2017). This approach necessarily entails a number of abstractions and cannot capture the richness of real-world indexical associations. This only shows that a multifaceted problem like indexicality needs to be approached from multiple angles. The main contribution of the work reported here is adding to the toolkit for studying the development of indexical associations.

In this study, we used a novel experimental paradigm and artificial stimuli to study how adults learn and process social–contextual cues to linguistic variation. We looked at the relationship between a nonlinguistic context and a linguistic pattern in a simple learning task. In the task, adult speakers learned to associate two specific morphophonological patterns with two contexts. This work is a conceptual replication of Rácz, Hay, and Pierrehumbert (2017)—following the definitions proposed by Marsden, Morgan-Short, Thompson, and Abugaber (2018). In Rácz et al. (2017), we demonstrated that such indexical context learning is possible in the laboratory and that it shows some notable differences from other types of category learning. In the current study, we went on to focus on a range of different types of social contexts and the precise mechanics of learning associations with these contexts. Our aim was to test whether certain types of context-language associations are learned faster than others and whether these are also generalized more easily to new linguistic and nonlinguistic contexts. Our task tested these questions by teaching an association between language and context.

Background Literature

Contextual Learning

People are able to rely on the context in learning tasks. They can create associations between linguistic or nonlinguistic categories and their context. Prior experience has an influence on what aspects of the context people focus on in a given learning task.

Considering context in a very broad sense, plenty of evidence has shown that a memory is easier to retrieve in the context in which it was established. In a classic study, Godden and Baddeley (1975) showed that words that were learned underwater were more accurately recalled underwater. In a related study, Hay, Podlubny, Drager, and McAuliffe (2017) showed location-specific effects on speech perception in the laboratory and in a car. Qian, Jaeger, and Aslin (2014) demonstrated in a nonlinguistic example of contextual learning that, in a “whack-a-mole” type game, players were faster at predicting the location of the mole if the location was probabilistically cued by moving background images to which the player was not overtly oriented. Lewicki, Hill, and Czyzewska (1992, p. 796) reviewed much earlier work on implicit learning. They argued for the sophistication of automatic learning “as compared with consciously controlled cognition, the non-conscious information-acquisition processes are not only much faster but are also structurally more sophisticated in that they are capable of efficient processing of multidimensional and interactive relations between variables." Evidence indicates that people transfer contextual cues to language tasks as well.

Work on language processing has shown that learned associations between context and language usage play an important role. Van Berkum, Van den Brink, Tesink, Kos, and Hagoort (2008), for example, looked at neural activity in speech comprehension using event related potentials. They found that listening to pragmatic violations that arise from contextually incongruous sentences (e.g., *I have a large tattoo on my back* spoken with an upper-class accent) result in neural activity that is comparable to semantic violations in sentences (e.g., *The Earth revolves around the trouble*). This suggested very early involvement of contextual information in sentence processing. Furthermore, within the set of incongruous sentences, they found additional differences in neural activity for sentences that were incongruous with a female or male speaker (e.g., *I like fishing on the weekend* for a female voice and *I hate having my period* for a male voice). The gender distinction in their stimuli provoked a stronger reaction in the participants than age and class distinctions did.

Molnar, Ibáñez-Molina, and Carreiras (2015) provided another example of the context feeding into language processing. They found that bilingual listeners were able to adapt to different interlocutors in spoken language processing by using contextual cues to language background provided by the interlocutors' identities. Brunellière and Soto-Faraco (2013) showed similar listener sensitivity to accent variation.

These examples support the efficiency of automatic learning. Selective attention also plays an important role in linguistic processing and contextual language learning. For instance, Leung and Williams (2012) showed that participants can implicitly (without awareness) attend to a grammatical agreement rule involving animacy but do not attend to one involving the relative size of two objects. They speculated that learner experience is vital in these contexts. One possible explanation for their results, they suggested, was that the critical variable driving implicit learning of form-meaning connections is not their availability in themselves. Rather, it is the availability of form-meaning connections to grammatical processes and representations, based on individuals' prior linguistic knowledge. Leung and Williams (2013) went on to demonstrate that learnability differs across learners with different language backgrounds. Speakers of Chinese learned a mapping between articles and a concept related to the Chinese classifier system, whereas speakers of English did not.

Sociolinguistic Variation and Context

Attention to detail in learning context-memory associations is reflected in the richness of these associations in language. Language-context associations provide an important starting point for sociolinguistics, and we can provide only a cursory overview of the ways in which they are relevant to language and society.

Language variation is linked to the social backdrop of language use in complex ways. The speaker and the addressee's positions in society, their relation to each other, and the context of their interaction all play a role in determining which linguistic variants are used. Social variables influence linguistic variation at all levels—from phonetic to syntactic variation. People rely on the social meaning of lexical items both in speech perception (Campbell-Kibler, 2011; Foulkes, Docherty, Khattab, & Yaeger-Dror, 2010; Giles, Taylor, & Bourhis, 1973; Hay, Nolan, & Drager, 2006; Jannedy, Weirich, & Brunner, 2011; MacFarlane & Stuart Smith, 2012; Niedzielski, 1999; Pharao, Maegaard, Møller, & Kristiansen, 2014) and speech production (Eckert, 2000; Foulkes & Docherty, 2006; Hay & Drager, 2007; Labov, 1972, 2001; Lawson, Scobbie, & Stuart-Smith, 2011; Milroy, 1980; Timmins, Tweedie, & Stuart-Smith, 2004; Trudgill, 1974). Speakers are able to keep track of the effect of context even at the word level (Hay et al., 2019; Pierrehumbert, 2016; Pierrehumbert, Beckman, & Ladd, 2000).

Users of different dialects will focus on and learn different linguistic details in interactions (Cohn, Ham, & Podesva, 1999), and their awareness of contextual information on all levels can be very imprecise and is often worse than assumed by even the speakers themselves (Preston, 1996). For instance, in an experiment by Clopper and Pisoni (2004), American English listeners were above chance in identifying the dialect region of American English speakers based on phonological differences alone, but their accuracy remained low.

In sociolinguistic variation, some social contexts are more relevant than others, echoing work on contextual learning. Some variables, like gender (Cheshire, 2002; Milroy & Milroy, 1993), age (Sankoff & Blondeau, 2007; Walker & Hay, 2011), and ethnicity (see, e.g., Johnson & Buttny, 1982) frequently show systematic influences on linguistic variation. Other types of group membership can be highly idiosyncratic and specific to a particular speech community (see, e.g., Gudmestad, 2012; Habick, 1991; Mendoza-Denton, 1996).

Sociolinguistic Learning

Of course, despite the affinity for learning context-language associations, people do not start with a perfect knowledge of sociolinguistic variation. They learn these associations along with the other denotative and structural aspects of language.

Learning the associations present between nonlinguistic contexts and linguistic patterns starts early (Foulkes et al., 2010; Smith, Durham, & Richards, 2013) and continues into adulthood, as evidenced by ongoing changes in the linguistic variation of individuals, mirroring changes in their communities (Harrington, Palethorpe, & Watson, 2000). The mechanisms through which we acquire knowledge about social variation are not fully understood. Its early appearance has been used to argue that the process is not distinct from the acquisition of denotative meaning but rather that denotative and social meaning both emerge from the same contextually and socially rich store of detailed linguistic memories (Chevrot & Foulkes, 2013; Pierrehumbert, 2006). Indeed, modern theories of the mental lexicon, that is, the storage of linguistic forms, tend to argue that nonlinguistic information (e.g., characteristics of speakers or the environment) and linguistic contextual information (e.g., distribution in a sentence) both play a crucial role in how forms are stored and processed, with effects on a range of phenomena from speech perception (Johnson, 1997) to priming (De Vaan, Schreuder, & Baayen, 2007).

Docherty, Langstrof, and Foulkes (2013) investigated the learnability of several types of sociophonetic associations using a methodological paradigm that involved passive exposure to words produced by two “tribes.” Across different experiments, the tribes differed in terms of the phonetic markers that distinguished them one from the other. In a subsequent test phase, participants listened to the same recordings to which they had been exposed and were asked to overtly label them as originating from Tribe 1 or Tribe 2, which they did with above-chance accuracy. Langstrof (2014) reported a number of further experiments using this paradigm. These studies focused on existing types of sociophonetic variation and restricted tests of pattern learning to words already encountered in training.

The result that is most relevant to the current work from Docherty, Langstrof, and Foulkes’s study was the demonstration that adult listeners do form associations between linguistic variants and social agents even after relatively little exposure. The strength of the association formed seems to vary across participants and be affected by the type of phonetic variation involved. In one experiment (Docherty et al., 2013), for example, socio-indexical variation involving consonants was more robustly learned than variation involving vowels. Although these studies showed that sociolinguistic learning is possible in the laboratory, the types of nonlinguistic contexts remained deliberately artificial. This was despite the existing assumption that some nonlinguistic contexts are easier to recognize and learn than others. For instance, Foulkes (2010) hypothesized that some types of social–contextual properties should be more readily transmitted and learnable than others due to the variable frequency with which properties had been relevant in individuals' past experience. This ties in with the observation that some nonlinguistic contexts are more strongly associated with sociolinguistic variation than others. Foulkes went on to identify speaker gender as one of the very earliest learned socio-indexical associations. Of course, gender is also marked in the grammar in most Indo-European languages.

Existing work has provided evidence that a linguistic association with gender is indeed learnable in the laboratory. Samara, Smith, Brown, and Wonnacott (2017) showed that both children and adults are able to associate a linguistic pattern with speaker identity in an experimental setting even if the association between pattern and speaker is variable. One of the main differences between their two speakers, Henry and Katie, was gender, indirectly providing experimental evidence for Foulkes' assumption. Expanding on this theme, Needle and Pierrehumbert (2018) went on to show that adult speakers of English pick up gendered associations of words and morphemes from the ambient language and can generalize these associations to complex pseudowords. Hay et al. (2019) showed that a word's associations with both gender and age can affect lexical access patterns.

Implicit to much of the work on sociolinguistic learning is that associations between language and a nonlinguistic context are generalized. Such associations are, for the most part, formed between context and linguistic categories and are not restricted to individual items. The notion of context (*woman*, *person from Yorkshire*) can be applied to unfamiliar conversation partners based on their speech. Hybrid models of language variation (Pierrehumbert, 2006) predict both instance-specific learning and generalization to more abstracted categories. As Pierrehumbert (2006) pointed out, however, there is much that is not understood about how this works in the social domain. The less populous samples available for some social categories (compared to phonological categories) may lead them to be less robustly learned. Few studies in the literature have directly probed the mechanics of generalizing learned socio-indexical associations. Sneller and Roberts (2018), for instance, showed that the adoption of new sociolinguistic variants in the laboratory hinged on the contextual associations of the variants, hinting at the joint role of context and generalization in sociolinguistic learning, but Sneller and Roberts did not address the specifics of within-context generalization directly.

In Rácz et al. (2017), we looked at the process of learning social meaning by comparing types of linguistic and nonlinguistic contexts. The focus of Rácz et al. (2017) was on the process of learning, contrasting learning in a strictly linguistic context (matching the vowel of the suffix to the vowel of the stem) to learning in a mixed context (matching the suffix to the conversation partner), and on the link between training and generalization strength. The results showed that learning and generalization are not robustly different in a linguistic context from learning and generalization in a mixed context. At the same time, instance-based generalizations were more important for a linguistic context, whereas nonlinguistic associations were treated more generally. Although it was true that some participants were able to rely on nonlinguistic associations for all the contexts discussed by Rácz et al. (2017), the number of successful learners clearly hinged on the type of context. In particular, participants were best at associations with a gender-based difference between conversation partners. In the current study, we further explored the varying strength of influence of nonlinguistic contexts. The 2017 study tested two nonlinguistic contexts, one based on conversation partner gender, and another based on conversation partner spatial orientation. The focus of the current study was variation in learning accuracy and the extent of generalization of learning across various other nonlinguistic contexts.

The Current Study

Previous research has demonstrated that adults are able to keep track of a large amount of very detailed nonlinguistic context in language processing (Needle & Pierrehumbert, 2018) and learning tasks (Docherty et al., 2013) and that they rely on prior experience to weigh contexts differently (Molnar et al., 2015) and to discard contexts that are irrelevant (Leung & Williams, 2012). These aspects of contextual learning are reflected in sociolinguistic variation, which uses a wide array of contexts, some of which are more readily associated with linguistic variation than are others. In the Background Literature, we showed that sociolinguistic learning is a lifelong process, one that can be studied in the laboratory. Several studies, including our own, have shown that participants can learn a gender-based contextual distinction in an experimental setting.

These results have brought new questions to the fore. For example, are gendered patterns learnable in the laboratory because gender of the speaker is highly salient, as demonstrated by its early acquisition (Ladegaard & Bleses, 2003), as well as by behavioral and neural evidence (Cheshire, 2002; Van Berkum et al., 2008)? As we discussed in the Background Literature section, speaker age and ethnicity are also prevalent variables in sociolinguistic variation. Is this because these nonlinguistic contexts are also eminently learnable? Is there a difference in learnability between gender, on the one hand, and age and ethnicity, on the other? Does this difference translate to how easily these contexts generalize?

This study aimed to address these questions couched in the broader frameworks of sociolinguistic variation and sociolinguistic learning. We wanted to further demonstrate that we can not only attest sociolinguistic learning in the laboratory but also test for various factors that may influence it. We also wanted to demonstrate that an artificial language task, despite the necessary simplifications that it entails, can add to the toolkit for studying language variation and change by shedding further light on its mechanics. This led us to our starting hypotheses:

1. A nonlinguistic contextual cue that is socially salient (e.g., conversation partner gender) is easier to learn than a socially irrelevant cue (e.g., conversation partner spatial orientation). Rácz et al. (2017) found some evidence for this. Within the set of socially salient cues, some (such as gender) are stronger and therefore easier to learn than others (such as age or ethnicity). In a setting where contextual cues compete for attention, a socially more salient cue is also harder to ignore.
2. Learning is followed by generalization: A pattern that is learned with a socially salient contextual cue is easier to generalize to new linguistic and nonlinguistic contexts than a nonsalient cue.

Method

Design

The backbone of our design was a simple artificial language learning task adapted from Rácz et al. (2017). It consisted of a training phase followed by a test phase. In training, a player took the role of a chicken that flew from roof to roof, going toward a destination (its nest). The chicken met various conversation partners, one in each trial. The conversation partner showed the chicken a prompt image (e.g., a jug) with an accompanying name (e.g., *fen*). The chicken had to respond to this prompt with a related named image. The response image was given, but the player had to pick the name from two options (e.g., *fenpel* and *fenfis*). A correct answer meant that the chicken could go on, an incorrect answer resulted in going back two trials. The player had to pick the correct answer to every question to finish training. For the code and an example of the training, see https://github.com/nzilbb/roofRunner.

The upper image in Figure 1 shows the layout of a training trial. The trials are entirely visual. On the left, is the chicken; on the right, the conversation partner, an adult female wearing a yellow top. A speech bubble ties the prompt to the conversation partner. The speech bubble is the image of a large gate with an accompanying name: *fen*. The response comes from the chicken. The image is a diminutive version of the gate, and the two possible responses are presented as two buttons: *fenpel* and *fenfis*. It is evident that the first part of each response is the word for gate (*fen*) and what varies is the suffix: -*pel* or -*fis*.

In the first trial, the player has to guess because no clue is given to the correct answer. If the correct answer is *fenpel* in this example, choosing it will take the player to the next trial. The second image in Figure 1 shows this second trial. The prompt image is now a mushroom (*rik*) and the response is the name of the diminutive mushroom, the two options being *rikpel* and *rikfis*. Again, the suffixes are clearly carried over from the previous trial. The conversation partner is different: The partner is now an adult male in a black t-shirt.

Choosing the wrong answer (*rikpel*, with the -*pel* suffix) sends the player back two trials. Choosing the correct answer (*rikfis*) takes the player to the next trial (Figure 1, third image). Here, a new named object appears as well as a new conversation partner image: This is clearly the female from the first trial, but shown sideways. The correct answer is -*pel* again. These three trials allow an observant player to figure out the rules of this training task: The woman prefers -*pel*, and the man prefers -*fis*. The lower panel of Figure 1 shows a subset of the prompt images used in the experiment.

The test phase followed the training phase. A test trial was similar to a training trial except that it occurred at night rather than during day, the chicken was no longer present, and the players received no feedback. The test phase included items and conversation partners familiar from the training phase as well as entirely novel ones, as Figure 2 shows. The aim of the test phase was to determine whether the pattern learned in the training phase carried over and whether it generalized to items and conversation partners not seen in training. In the example above, a previously unseen item and a male partner not seen in the training phase should still have led players to pick the male suffix, -*fis*.

Rácz et al. (2017) employed a basic design that used the four conversation partners seen in Figure 3. In this design, all players saw the same four conversation partner images in the same setup. The difference was how these images were grouped. In the so-called gender condition outlined above, the correct answer (-*pel* or -*fis*) depended on whether the conversation partner was a woman or a man. The way that they faced was irrelevant. In the corresponding view condition, the correct answer depended on which way the conversation partner

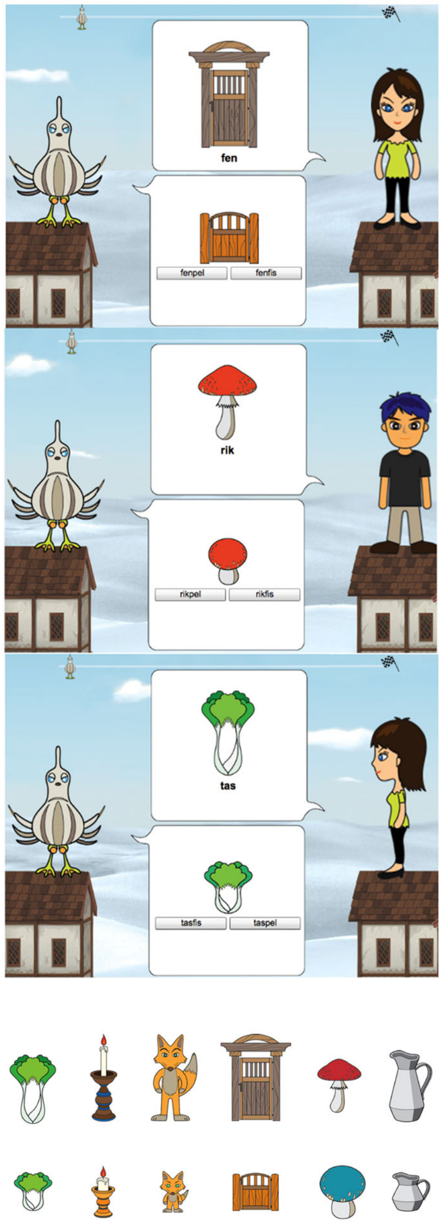

Figure 1 The layout of the training phase of the artificial language learning task (above). A sample of prompt and diminutive images (below). [Color figure can be viewed at wileyonlinelibrary.com]

Figure 2 The layout of the test phase of the artificial language learning task. [Color figure can be viewed at wileyonlinelibrary.com]

faced; people facing front preferred -*pel*, but conversation partners facing sideways preferred -*fis*, for instance. A player was randomly assigned either to the gender or to the view condition.

In Rácz et al. (2017), we found that the gender distinction was learned and generalized much more easily than the view distinction. This learning extended to new items and new conversation partners (e.g., from women to girls and men to boys). In the current study, we expanded the set of conversation partners in two ways. The gender and view distinctions are also compared to a distinction in conversation partner's age and ethnicity in the training phase. Categories are extended in new ways in the test phase (e.g., from female conversation partners to male conversation partners). We discuss this in detail below.

Materials

Players always saw four conversation partners in the training phase and four additional ones in the test phase. For an example, see https://github.com/nzilbb/roofRunner. These partners varied from player to player. Conversation

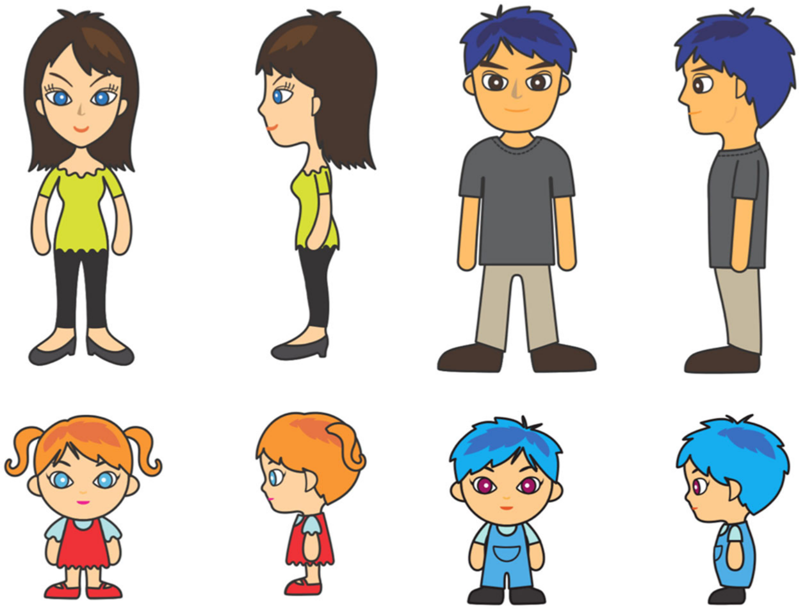

Figure 3 The four conversation partners in the Rácz et al. (2017) study. They differ in gender, age, and view (spatial orientation), of which gender and view were tested. [Color figure can be viewed at wileyonlinelibrary.com]

partners were drawn from a complete set seen in Figure 4. They can be grouped in four ways: by gender, age, ethnicity, and view. Not all combinations exist. A player saw two pairs of partners (four in total) in the training phase and an additional two pairs in the test phase.

The two pairs in the training phase could be grouped in two ways. One of these determined the suffix choice—this we call the main cue. The other one should be ignored by the player—this we call the competitor cue.

The four pairs in the test phase could be grouped in an additional way: Four of a kind were familiar from training, and four of a kind were new. The new conversation partners introduced a new dimension of contrast, for example, if training was with adults, the new partners might all be children.

In the example from the previous section, the main cue was gender, which had determined the correct answer in training; the competitor cue was view, the partner's spatial orientation, which was to be ignored. The test phase introduced new partners drawn from the pool of images unused in the training. For example,

Figure 4 The eight conversation partners in the current study. They differ in gender, age, ethnicity, and view (spatial orientation). [Color figure can be viewed at wileyonlinelibrary.com]

if partners in the training phase had been adults, the new, unfamiliar partners in the test phase could be children.

So far, we have discussed conceptual similarities across our conversation partner images (such as same gender, age). However, these images were also perceptually similar to each other in shape and color, and this partially overlapped with conceptual similarities. We were exclusively interested in conceptual similarity, and we used post hoc checks to control for the effect of perceptual similarity.

Participants encountered six items in the training phase, and they encountered these same six items along with six additional items in the test phase. We distributed the 12 item images randomly across the test phase and the training phase for each participant. We also picked suffixes and assigned stems to images randomly for each participant.

We built item names in the task from an artificial language. Because the focus of the task was social association, the artificial language itself was deliberately simple. For each participant, we drew syllables randomly from a finite syllabary that contained the following 14 items:

fek; rul; rik; wan; wuk; fen; fal; wun; pel; tas; ril; fis; tol; tos.

We used two syllables as suffixes, and we randomly assigned the remaining 12 syllables as names to the 12 items used in the training and test phases (i.e., six syllables in the training phases to which we added six more syllables for the test phase).

The syllabary reflected the following design principles:

- the syllables should be distinctive;
- they should consist of a small set of frequent letters;
- they should be easy to pronounce for our participants, who were American English speakers; and
- the consonant clusters in the two-syllable words should cue English word boundaries in a uniform manner.

Our aim was to provide an optimal set that balanced these considerations.

We used the diminutive as the contextually cued morphology because the diminutive is a common, iconic pattern that is easy to interpret visually. However, the diminutive form is highly variable in English, and it has strong associations with gender in many languages (Jurafsky, 1993). In order to make our findings more robust, we repeated two conditions using plural instead of diminutive as the morphological category. In these repeated conditions, participants performed the same task with plural images rather than with diminutive ones. The words were similar, the implied meaning different.

In the diminutive condition, the representation of the target item was a smaller, exaggerated, cuter version of the large item (see Figure 1 for a mushroom and a tiny mushroom). In the plural condition, the representation of the target item was a picture of three of the target items, normally scaled instead of one diminutive version.

Participants

The experiment was hosted on Amazon Mechanical Turk. The platform has been used successfully (albeit with caveats) in behavioral research (Crump, McDonnell, & Gureckis, 2013) and provides a participant pool that is more representative than typical convenience samples (Berinsky, Huber, & Lenz, 2012).

We ran participants in three large batches, each several weeks apart in 2014 (see Rácz et al., 2017) and 2015. Participants were paid $3 US when they had completed the task. We restricted participants to those with IP addresses in the United States, and participants had to be native speakers of English. We collected participant background information using a pre-task questionnaire.

A total of 474 participants took part in the experiment. We removed 11 people based on test-phase performance: In the test phase, these participants always clicked on either the first or the second button. This left 463 participants. We also filtered for participants who finished the task but had taken a disproportionate amount of time in training.

Based on timestamps recorded by Amazon servers, the mean length of the training phase was 5.28 minutes ($SD = 2.40$). The fastest participant finished in 1.60 minutes, the slowest (after filtering) in 18.93 minutes.

The duration of individual trials, however, provided an unreliable metric because participants played the game on their own computers and not in a laboratory setting. A participant who became distracted or stood up to make a cup of tea took longer to finish a trial, much the same way as did a participant who had difficulty making a choice. Instead, we used the number of trials needed by a participant to finish the training phase as our main indicator of participant speed. The distribution of participant trial counts had a minimum of 24 trials (six items seen with four conversation partners, and a participant responding correctly to every combination) and a long tail.

Participant sample sizes varied across the conditions. This was partly due to variability in exclusion rates and partly to shifting experimental protocols. We aimed at a minimum sample size of 30 participants per condition. We

Table 1 Combinations tested in the experiment

Variable	Age	Ethnicity	Gender	View
Age		√	√	√
Ethnicity	√		√∗	
Gender	√	√∗		√
View	√		√	

Note. Figure 4 is helpful in interpreting Table 1. A check mark means that the combination was tested in the experiment. Some combinations were not possible (a cue could not be a main cue and a competitor cue at the same time), and others were left out to streamline the design. All combinations were tested with the diminutive category. Combinations marked with an asterisk were also tested with the plural category. For any combination of two cues, a third cue was used to introduce new conversation partners in the test phase. This additional third was never the view cue to help the interpretation that new conversation partners were different individuals.

took additional steps to make sure that our sampling did not affect the results (discussed in the Results section).

Because we assumed that the training phase was longer in certain across-participant conditions than in others and because we wanted outlier thresholds to reflect this, we took every condition separately and then removed the slowest 2.5% of the participants based on a threshold count of the number of training trials completed (within conditions). This left us with 435 participants. Using simulations, we determined that a participant playing by chance would finish the training phase in about 518 trials. None of our participants was this slow.

Procedure

The training phase consisted of six images with four conversation partners (24 trials), whereas the test phase consisted of 12 images (six unfamiliar and six familiar) with eight conversation partners (four familiar and four unfamiliar; 96 trials). Each participant saw 120 unique trials in total.

The task had three across-participant variables, the main training cue and the competitor cue (gender, age, ethnicity, or view) as well as the type of morphological pattern (diminutive or plural). It had two within-participant variables, specific to the test phase: whether the conversation partner or the target item was familiar, that is, whether the specific conversation partner or the target had been encountered in training. Tables 1 and 2 provide an overview of the experiment.

Table 1 shows the existing main cue/competitor cue combinations in the experiment. These were the effective across-participant conditions.

Table 2 Summary statistics for the test phase

Main cue (competitor cue)	Pattern	*n*	Ratio$_{men}$	M_{age} (years)	*SD*
Age (ethnicity)	Diminutive	30	.53	30.59	9.97
Age (gender)	Diminutive	36	.34	30.77	6.88
Age (view)	Diminutive	49	.37	34.55	11.58
Ethnicity (age)	Diminutive	30	.38	33.28	13.40
Ethnicity (gender)	Diminutive	26	.64	32.40	12.50
Ethnicity (gender)	Plural	35	.54	30.54	8.46
Gender (age)	Diminutive	37	.59	33.00	10.28
Gender (ethnicity)	Diminutive	26	.54	29.54	8.38
Gender (ethnicity)	Plural	31	.39	32.45	7.83
Gender (view)	Diminutive	47	.38	33.74	11.58
View (age)	Diminutive	40	.40	32.90	9.54
View (gender)	Diminutive	48	.48	30.47	9.52

Table 2 shows the participant counts and summary demographics across the across-participant conditions.

Data Analysis

We report results primarily from the test phase. We used participants' training performance as a predictor of their accuracy in the test phase. We report the estimates of a main model fit on the diminutive data only and provide a series of secondary models to test the robustness of the test-phase results.

We used the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) implemented in R (R Core Team, 2018) for model fitting and ggplot (Wickham, 2016) for plots. In Rácz et al. (2017), we reported results for the gender (view) and view (gender) conditions. We report these here again, analyzed in combination with the other conditions. We fit Model 1 on test-phase data for the diminutive. We compared diminutive and plural data in Model 2. We tested the effect of perceptual versus conceptual distance between conversation partner images in the test phase in Model 3. In Model 4, we refit Model 1 while resampling the participants.

For the main model (Model 1), we first specified a model with all main terms and no interaction terms. We then tested all relevant interactions by adding them one by one and using model comparison to determine whether this leads to a loss in the amount of variation explained by the model. We tested the robustness of individual main terms in a similar way; we removed them one by one and used model comparison.

A number of criteria exist for model comparison. We relied on chi square goodness-of-fit tests to select the best model. We have outlined the model fitting process in detail in Appendix S1 in the online Supporting Information. Below we report the model with all main terms and robust interactions. We included the main terms because they express aspects of the experimental design. The interactions were more exploratory and post hoc, and so we have included only robust ones in the reported model.

For Model 1, Model 2, and Model 3, we have reported the results of goodness-of-fit tests and have provided more details in Appendix S1. For Model 4, we ran a Monte Carlo simulation and have provided results with an error threshold.

Results

We had hypothesized that participant accuracy in the test phase would vary for (a) main cues compared to competitor cues and (b) for previously seen items/conversation partners compared to those present only in the test phase. In Rácz et al. (2017), we investigated two contextual cues, meaning that the main cue always implied the competitor cue. In the current study, main cue and competitor cue were independent, and we considered them separately.

Primary Analyses to Examine the Learning of Contextual Associations

We fit Model 1 to test these hypotheses for the diminutive pattern using multilevel binomial generalized linear regression. The outcome was whether the participants picked the correct names for the diminutive items in the test phase. (The correct name at test is the name that was preferred for that item by the conversation partner group encountered during training.) The predictors were the main cue type, the competitor cue type, whether the item was familiar from training, and whether the conversation partner was familiar from training. We tested all non-rank–deficient interactions of these predictors. In addition, we included the participants' training trial count as a predictor. The model also had a participant random intercept. Because we had randomized item images and names, we did not include an item grouping variable.

Figure 5 shows the estimates for the best model for the results of the test data (Model 1, ID 2; see Table 3 in the appendix) with Wald 95% confidence intervals. (For the values, please see Appendix S1.) The Wald 95% confidence intervals capture the certainty of estimates: Where the interval excludes 0, we can be 95% certain that the true difference is nonzero.

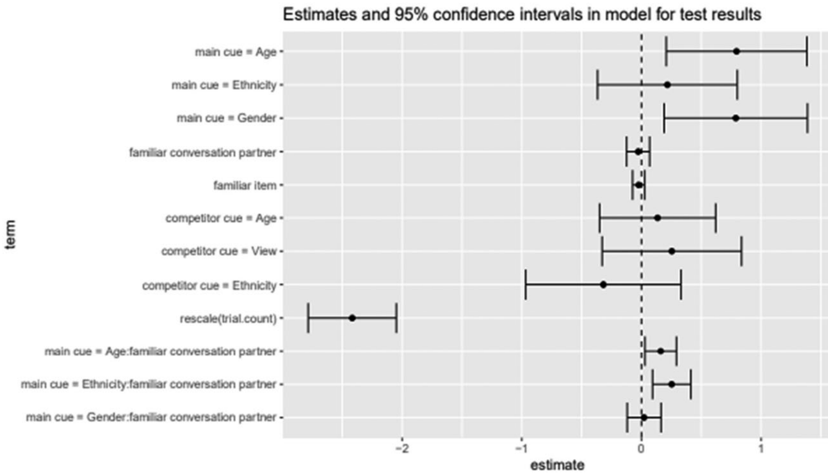

Figure 5 Estimates of Model 1 with Wald 95% confidence intervals for diminutive test data.

To explain our results, we first present the term estimates, then use visualizations of our data to expound on the relevant patterns, and finally provide a summary of how these patterns relate to our hypotheses.

Figure 5 shows that accurate responses to a test-phase trial varied across main cue type. Participants were more accurate with the gender (est = 0.79, 95% CI [0.19, 1.39]) and age cues (est = 0.79, 95% CI [0.21, 1.38]) compared to the view (intercept) and ethnicity cues (est = 0.22, 95% CI [−0.37, 0.8]). This was mediated by familiarity with the conversation partner. This interaction was relevant for the age cue (est = 0.16, 95% CI [0.03, 0.29]), and the ethnicity cue (est = 0.25, 95% CI [0.09, 0.41]). Participants who finished training in more trials were also less accurate in the test phase (est = −2.42, 95% CI [−2.79, −2.05]). The competitor cue, and familiarity with items, made no robust difference in determining participant accuracy. (The relevance of these terms was tested using goodness-of-fit tests that are reported in Appendix S1.)

Main cue is an unordered variable, and the implications of term estimates can be relatively hard to interpret. Visualizations of the test data shed more light on participant behavior. Figure 6 shows the main effect of cue type. The figure displays participants' mean accuracy in the test phase across main cue type (upper panel) or competitor cue type (lower panel). The individual points are participant means. The violins show their distributions across cue types, and the black points show the means of these distributions.

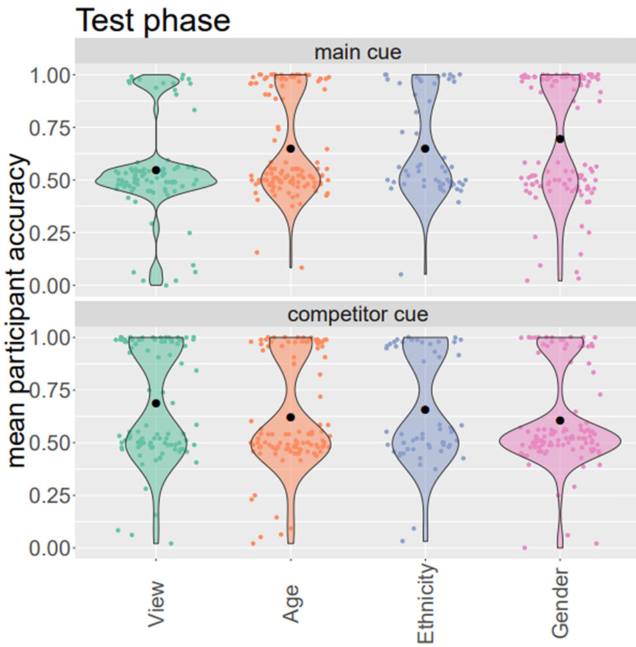

Figure 6 Participant means across main cue (upper) or competitor cue type (lower) in the test phase. Colored dots (dots around the violins) are jittered participant means; black dots are means for cue types. [Color figure can be viewed at wileyonlinelibrary.com]

The first interesting aspect of participant distributions is that they vary across main cue type. For main cues, participants who had to learn the view cue in the training phase were the least accurate, overall, in the test phase. Participants learning the gender cue were the most accurate, with age and ethnicity in between. For competitor cues, participants who had to ignore the view cue were the most accurate, but participants who had to ignore the gender cue were the least accurate. However, variation across competitor cues was much less pronounced. Model 1 lends support to meaningful differences in test accuracy across main cue type, though it does not warrant an absolute order of difficulty. We did not find such support for differences across competitor cue type.

The second interesting aspect of these distributions is that they are predominantly bimodal. In each distribution, a group of participants was clustered round .50 (chance level), and another group was close to 1 (ceiling). It is likely that participants who understood the rule that they had to learn in the training

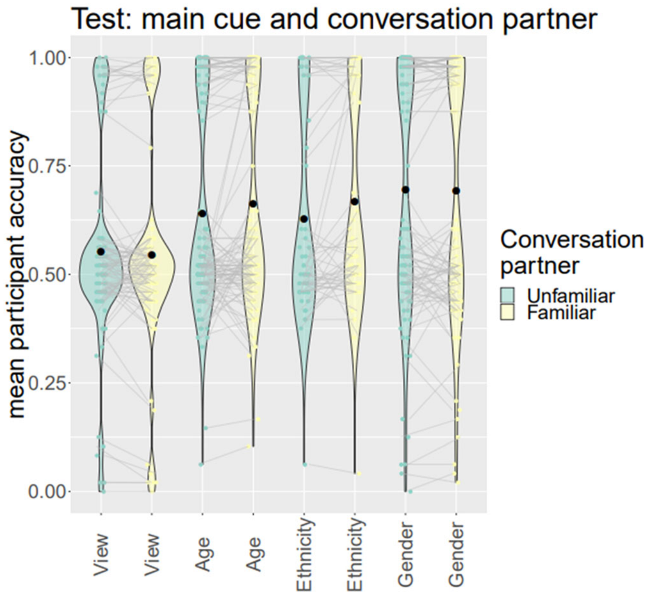

Figure 7 Participant means across main cue and familiarity with the conversation partner in the test phase. The figure breaks down the upper panel of Figure 6. Each violin is split into two: in each pair, the one on the left (green) for unfamiliar and the one on the right (yellow) for familiar conversation partners in the test phase. Each grey line represents one participant. [Color figure can be viewed at wileyonlinelibrary.com]

phase were the ones who were very accurate in the test phase, but those that kept guessing and passed the training phase by rote learning the correct answers were the ones who were also guessing in the test phase. It is, in fact, the proportion of these two sets across main cue type that shifted the overall means: More participants figured out the gender rule than did the view rule. This implied that participant success in the test phase correlated with participant success in training. Indeed, the training trial count was a significant main effect in the model. The bottom line is that participants were most accurate with the gender cue and least accurate with the view cue.

Model 1 indicates that participant accuracy across main cues was mediated by familiarity with the conversation partner, as Figure 7 illustrates. This figure replicates the upper panel of Figure 6, split according to whether the conversation partner was familiar from training. The grey lines connect a given participant's average for unfamiliar trials and familiar trials. These are the data underpinning the significant interaction retained in the model. For view and gender, familiarity with the conversation partner made no difference to participant accuracy. Participant means move up and down in the split to some extent, but no clear pattern emerges for the role of familiarity with partner when the main cue was view or gender. In contrast, for both main cues age and especially ethnicity, some participants were a lot more accurate in test trials with conversation partners who were familiar to them from training. This was despite the fact that new conversation partners shared the same grouping characteristics as the familiar partners (i.e., they were also children or adults).

To us, this is the main result of the experiment. Participants who had to learn a contextual distinction that was not supported by prior knowledge (i.e., the view cue) mostly kept guessing in the test phase overall, hence they showed no improvement with familiar conversation partners. For participants who learned the gender distinction in training, generalization was already so complete that they were at ceiling accuracy with both old and new conversation partners.

The two categories in between were eminently learnable but generalization to new conversation partners was not straightforward—participants benefited from familiarity. This challenges the interpretation that we alluded to above, namely, that participants either learned the context-pattern association or not. This seems to be true for the gender cue, but not for age and ethnicity.

In summary, a larger proportion of participants was successful at learning meaningful contextual associations (gender, age, ethnicity) than a nonmeaningful association (view). Gender, in particular, was the easiest for participants to learn out of the three associations. Gender-based associations were generalized straightforwardly to new conversation partners, but those based on age and ethnicity were also generalized, but to a lesser degree. This indicated that a robustly learnable distinction was also robust to generalization.

Secondary Models to Test the Robustness of Evidence for Learning Meaningful Contextual Associations

In order to assess the robustness of these results, we fit a number of secondary models to test potential confounding variables. First, we examined the experimental conditions for which we had collected both diminutive and plural data. We fit Model 2 on cue types tested with both the diminutive and the plural patterns: ethnicity/gender (here, gender was either the main cue, with ethnicity as the competitor, or the other way round). We fit the model once with pattern type as a predictor and once without it and used a χ^2 goodness-of-fit test for model selection. We found that pattern type did not explain more variation in the data, $\chi^2(df = 1) = 0.01, p = .94$.

Second, we verified that the results were not artefacts of the differences in the visual similarities of our images. This important post hoc check concerned the extent to which perceptual similarity between conversation partner images affected participant behavior. Though these two are intertwined, we wanted participants to react to conceptual similarities (e.g., same gender) without the interference of perceptual similarities (e.g., same height). We used a signal processing metric (Levenshtein distance between the images) as a measure of visual differences because reliance on human raters would necessarily have invoked social-conceptual distances as well in determining visual distance. In this post hoc test, perceptual distance is an unsystematic variable and we want to see if it has any systematic predictive power over the associations made.

We calculated the Levenshtein distance for all conversation partner image pairs in our training data. We used the Image Processing Toolbox of MATLAB (MathWorks, Inc., 2016). We then matched these distances to each individual participant and aggregated over main cue type and competitor cue type. This gave us an aggregated main cue distance and a competitor cue distance expressing the perceptual difference in the category on which the participant was trained versus the category that the participant needed to ignore. If perceptual distance was relevant, participants should have had higher accuracy in learning associations with more perceptually contrastive conversation partner categories.

We fit Model 3, a multilevel binomial generalized linear regression model, on the test data, predicting an accurate response in a test trial based on the perceptual image distance between the image pairs in the main and the competitor groupings in training. It also included the image pairs themselves as grouping variables.

We compared the model to the best fit of Model 1—excluding training trial count—to see which one explained more variation in the test data: the model that relied on perceptual distance or the one that relied on conceptual distance. We found that the model that relied on conceptual distance gave a much better fit than the one that relied on perceptual distance, $\chi^2(df = 8) = 173.55$, $p < .001$.

Finally, we verified that the unequal sample sizes were not responsible for our key result (i.e., that accuracy with main cues was mediated by familiarity with the conversation partner for the age and ethnicity cues, though not for the view and gender cues). Our fourth model was a replication of Model 1 with resampled participant sets. Sample size varied across conditions. In order to make sure that our results were not contingent on sample size variation, we resampled the data 100 times, sampling the same number of participants in all conditions. This number was the number of participants in the smallest sample, ethnicity (with gender as the competitor cue), which was 26. We fit a multilevel model with the interaction of main cue and familiarity with the conversation partner for each resampled data set. Using a z value of 1.80 as a cutoff, we found that the crucial interaction (i.e., of main cue with familiarity with the partner) remained robust with the main cue age in 38/100 models and with main cue ethnicity in 96/100 models. The interaction of the main cues view or gender with familiarity with conversation partner was robust in 0/100 models. How does this relate to the Model 1 fit on all the data? There, we saw an interaction with the age cue and the ethnicity cue, but not with the gender cue or the view cue. We did not retrieve an interaction for gender and view in the sampling iterations, which is expected, since we found none in Model 1. We did recover an interaction for ethnicity, and, to a lesser extent, for age. This indicated that the interaction specified in the best fit of Model 1 was not an artefact of the variation in sample sizes.

To sum up the results of the experiment, main cues showed stratification in participant accuracy—players struggled to learn an association with the view cue, but they were likely to easily learn, and to reliably generalize, an association with gender. Associations with age and with ethnicity were learned moderately well, and, in our data, they showed evidence of some instance-specific learning, with the process of generalization still underway.

Discussion

The task of learning an association between a nonlinguistic context and an allomorphy pattern was a hard but not impossible one for our participants. Participant accuracy reflected real-life sociolinguistic knowledge brought into the experiment. Training took longer and test accuracy was lower with a less socially relevant cue compared to training and test accuracy for socially relevant cues. Within our small set of socially relevant contextual cues, participant behavior in the test phase indicated that the context-pattern association was initially rote learned and then generalized to new contexts (new conversation partners). This generalization was easier with a socially more robust cue.

In effect, one can think of training with any of the socially relevant cues (gender, age, and ethnicity) as learning an association between a suffix and individuals: Two individuals will prefer suffix A while the other two will prefer suffix B. In the test phase, the participants had to recognize that the new individuals whom they now encountered shared characteristics with the ones whom they had previously met in training: If the two children preferred suffix A in training, all children would prefer it in the test phase, too. For our participants, this particular generalization was harder than recognizing that, if the two female characters preferred suffix A in training, all female characters would prefer it in the test phase.

It is interesting that the competitor cue did not robustly affect participants' accuracy. We speculate that this was because the paradigm only reinforced the main cue, which could have had an unbalancing effect on cue competition by focusing attention. Alternatively, it is possible that participants who focused on the competitor cue ended up making many mistakes in training and getting feedback that was not informative to them for building a system of associations. Based on work on error and corrective feedback, this would set up a situation where very little learning takes place (Metcalfe, 2017).

On the whole, the task demonstrated the use of prior nonlinguistic knowledge of the social differences that are commonly signaled by linguistic differences, matching work discussed in the Background Literature section. Spatial orientation, which could play a role in resolving deixis, but has no social salience, was the hardest to associate with a linguistic pattern in this task. This remained true despite the design's apparent simplicity, relying on a small artificial language and exaggerated cartoon representations of extant social constructions. It is possible that the nature of the linguistic target in this study (i.e., nouns) made it less likely that a spatial orientation cue was salient to participants, given that cues for resolving deixis might be more likely to be expressed on different parts of speech (such as subject and object pronouns).

In terms of providing methodological insight, our task showed how adult sociolinguistic learning can be dissected using an artificial language task. Sociolinguistic learning can be seen as a process that starts as an association of a linguistic pattern with a specific nonlinguistic context that is gradually generalized to other, similar contexts. Of course, how this generalization unfolds in real life and its possible limits were beyond the scope of this study.

The task design made it clear that allomorph selection is a response to the conversation partner and that incorrect responses impede success. Although the primary aim was to render the task as straightforward as possible, this setup also has real-life analogues. For example, in French, incorrect marking of the gender of an adjective can potentially lead to ambiguity or poor comprehension—compare *Je suis heureux* and *Je suis heureuse* ("I'm glad" masc/fem).

The task layout might have had an effect on the results. The task was entirely visual and responses were in a forced-choice format. An open format would likely have resulted in a different pattern of responses, but it would have also required the participants to effectively memorize the entire syllabary, rendering an already challenging task even more difficult.

With this in mind, we return to our starting hypotheses: The results presented here indicate that a socially salient cue is more learnable than an irrelevant cue and that certain salient cues are easier to learn than others. In addition, the salience of a cue plays into the extent to which the cue can be generalized (from familiar conversation partners to new ones). These results have expanded on our Rácz et al. (2017) study that showed that it is possible to study contextual language learning using artificial language methods. We expanded on this study by broadening the range of contexts that we investigated and found that their real-life prevalence manifests in how participants learn and generalize them.

Docherty et al. (2013) and Leung and Williams (2012) have shown that different types of linguistic variation are differently learnable in a way that is linked to prior experience. Here, we have shown that in the learning of socio-contextual meaning, different social variables also fare differently. We thus have provided experimental evidence supporting the hypothesis put forward by Foulkes (2010) that some types of indexical properties should be more readily transmitted and learnable than others. Foulkes identified interlocutor gender as one of the very earliest learned socio-indexical associations, and in our experimental paradigm, we have shown that—of the contextual variables that we tested—gender is the most easily attended to and learned by adults. (Experimental paradigms that cover contextual-learning are generally too simple to consider the sex–gender distinction and its implications for the structure of social knowledge.)

Our results are also in line with those of Samara et al. (2017), who showed that both adults and children can learn a gender-based association in an experimental setting, and those of Needle and Pierrehumbert (2018), who showed that the gendered associations of suffixes for American English speakers carry over to pseudowords. In our task, both the stems and the suffixes were pseudowords, and the gendered association was established during the task. Learning, at least within the task, still took place.

The causal mechanisms underpinning this result are hard to disentangle. On the one hand, one might argue that this shows the importance of prior experience in (potentially implicit) socio-contextual learning. Conversation partner gender is very likely among the most frequent variables our participants had encountered in terms of conditioning linguistic variation. Gender is also marked explicitly in parts of the English pronoun system. On the other hand, one might use these results to argue for the overall high salience of gender as a social category—a variable that might then itself lead to increased socio-contextual learning in the world and heightened transmission of gendered associations relative to other types of socio-contextual variation. Indeed, it is possible that both of these interpretations contain some truth and that these serve to reinforce each other.

Limitations and Future Directions

These results accentuate the need for a more nuanced understanding of mechanisms of linguistic variation and transmission. Much of the literature has documented how social variables are reflected and constructed through variation in speech, and much speculation and modeling has related to how these associations emerge and are transmitted (see Foulkes & Hay, 2015; Sneller & Roberts, 2018). A missing piece in this literature, however, is that not all social variables are equal in terms of how much people attend to them and how much they store them when processing and learning language. Salient social groupings are important not only in influencing the nuanced ways in which people produce and construct language but are also differentially implicated in the very information that they store when they encounter words in context. We have shown this with very crude groupings of gender, age, and ethnicity. Needless to say, in real language variation in the real world, much more subtle community-specific and individual-specific social variables are at play (Eckert, 2000).

Our results also point to interactions between the salience of a nonlinguistic contextual cue and the learning process itself because our participants found it easier to generalize a more salient context. The present study cannot account for the entire process of learning sociolinguistic variation, particularly because our participants were all over 18 years old and our design used one-to-one correspondences between context and patterns (women always used *-ril*, etc.). Consistency of input has a huge effect on learning associations, and child and adult learners react to input variability very differently (Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009). Although specific instances of linguistic address, for example, can behave categorically, such consistency is extremely rare in sociolinguistic variation in general. Even in purported cases of completely deterministic sociolinguistic patterns, such as the gendered languages described in the ethnographic literature, actual practice is more multifaceted, with playful and metalinguistic uses present (see, e.g., Trechter, 1995). In American English, the native language of our participants, gender is thought to be very rarely associated categorically with social language use (Eckert & McConnell-Ginet, 1992).

In addition, the closely related results of Samara et al. (2017) showed that both adults and children are able to generalize linguistic cues of speaker identity even if these cues are only probabilistically (rather than categorically) associated with the nonlinguistic context. It is, therefore, all the more remarkable that we saw learning and generalization vary considerably as a function of the type of social context in our study despite the absolute categorical consistency of linguistic input-context relations.

A further consideration is that a category is learned more robustly when information is distributed across a larger number of contexts, an aspect of contextual learning that we did not address in this study (Maye & Weiss, 2003; Maye, Werker, & Gerken, 2002, though see Atkinson, Kirby, & Smith, 2015). In Rácz et al., 2017, we showed that training with 18 (instead of six, as in the current study) items improved participant accuracy in the test phase. Nevertheless, in the current study, although the number of nonce words and the number of nonlinguistic context types were relatively low, differences in learning accuracy still emerged.

Our work used an artificial language to investigate the preexistence of social categories, whereas such categories are, in reality, probabilistically associated with existing, complex linguistic patterns over time. However, nonce words and artificial languages have been used with much success in psycholinguistics to investigate learning in children and adults since Berko's (1958) classic work, and artificial language paradigms have been valuable tools of investigating complex linguistic phenomena in a controlled environment (see Roberts, 2017; Scott-Phillips & Kirby, 2010).

Conclusion

Social meaning, its relationship with other aspects of meaning (such as reference), its reliance on general cognitive mechanisms, and how it is mediated by variation all constitute complex problems, problems that can only be addressed using a combination of experimental and field methods. We hope to have shown that a relatively simple paradigm can provide insights into the process of associating cues in the environment with linguistic cues; evidence that would be harder to gain from complex realistic data. However, such a paradigm cannot, in itself, address all questions regarding social meaning and how it is learned. This work thus marks just the beginnings of understanding how intersecting social variables are implicated in and impact socio-contextual learning.

Final revised version accepted 12 December 2019

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at https://doi.org/10.5281/zenodo.3519395 and https://github.com/nzilbb/roofRunner. All proprietary materials have been precisely identified in the manuscript.

References

Atkinson, M., Kirby, S., & Smith, K. (2015). Speaker input variability does not explain why larger populations have simpler languages. *PloS One*, *10*(6), e0129463. https://doi.org/10.1371/journal.pone.0129463

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351–368. https://doi.org/10.1093/pan/mpr057

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*, 150–177.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 185–207. https://doi.org/10.1109/tpami.2012.89

Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and Language*, *125*, 82–93. https://doi.org/10.1016/j.bandl.2013.01.007

Bucholtz, M. (1999). "Why be normal?": Language and identity practices in a community of nerd girls. *Language in Society*, *28*, 203–223. https://doi.org/10.1017/s0047404599002043

Campbell-Kibler, K. (2011). Intersecting variables and perceived sexual orientation in men. *American Speech*, *86*, 52–68. https://doi.org/10.1215/00031283-1277510

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Science*, *10*, 335–344. https://doi.org/10.1016/j.tics.2006.05.006

Cheshire, J. (2002). Sex and gender in variationist research. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *Handbook of language variation and change* (pp. 423–443). Oxford, UK: Blackwell. https://doi.org/10.1002/9780470756591.ch17

Chevrot, J.-P., & Foulkes, P. (2013). Introduction: Language acquisition and sociolinguistic variation. *Linguistics*, *51*, 251–254. https://doi.org/10.1515/ling-2013-0010

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

Cohn, A. C., Ham, W. H., & Podesva, R. J. (1999). The phonetic realization of singleton-geminate contrasts in three languages of Indonesia. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 587–590). Berkeley, CA: ICPhS.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

De Vaan, L., Schreuder, R., & Baayen, R. H. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon*, *2*, 1–23. https://doi.org/10.1075/ml.2.1.02vaa

Docherty, G. J., Langstrof, C., & Foulkes, P. (2013). Listener evaluation of sociophonetic variability: Probing constraints and capabilities. *Linguistics*, *51*, 355–380. https://doi.org/10.1515/ling-2013-0014

Eckert, P. (2000). *Linguistic variation as social practice: The linguistic construction of identity in Belten High*. Hoboken, NJ: Wiley-Blackwell.

Eckert, P., & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology*, *21*, 461–488. https://doi.org/10.1146/annurev.anthro.21.1.461

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology*, *1*, 5–39. https://doi.org/10.1515/labphon.2010.003

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

Foulkes, P., Docherty, G., Khattab, G., & Yaeger-Dror, M. (2010). Sound judgments: Perception of indexical features in children's speech. In D. R. Preston & N. Niedzielski (Eds.), *A reader in sociophonetics* (pp. 327–356). New York, NY: De Gruyter Mouton. https://doi.org/10.1515/9781934078068.3.327

Foulkes, P., & Hay, J. (2015). The emergence of sociophonetic structure. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 292–313). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118346136.ch13

Giles, H., Taylor, D. M., & Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: Some Canadian data. *Language in Society*, *2*, 177–192. https://doi.org/10.1017/s0047404500000701

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*, 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Gudmestad, A. (2012). Acquiring a variable structure: An interlanguage analysis of second language mood use in Spanish. *Language Learning*, *62*, 373–402. https://doi.org/10.1111/j.1467-9922.2012.00696.x

Habick, T. (1991). Burnouts versus rednecks: Effects of group membership on the phonemic system. In P. Eckert (Ed.), *New ways of analyzing sound change* (pp. 185). 212). San Diego, CA: Academic.

Harrington, J., Palethorpe, S., & Watson, C. I. (2000). Does the queen speak the Queen's English? *Nature*, *408*, 927–928. Retrieved from http://doi.org/10.1038/35050160

Hay, J. (2018). Sociophonetics: The role of words, the role of context, and the role of words in context. *Topics in Cognitive Science*, *10*, 696–706. https://doi.org/10.1111/tops.12326

Hay, J., & Drager, K. (2007). Sociophonetics. *Annual Review of Anthropology*, *36*, 89–103. https://doi.org/10.1146/annurev.anthro.34.081804.120633

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*, 865–892. https://doi.org/10.1515/ling.2010.027

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *Linguistic Review*, *23*, 351–379. https://doi.org/10.1515/tlr.2006.014

Hay, J., Podlubny, R., Drager, K., & McAuliffe, M. (2017). Car-talk: Location-specific speech production and perception. *Journal of Phonetics*, *65*, 94–109. https://doi.org/10.1016/j.wocn.2017.06.005

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

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151–195. https://doi.org/10.1080/15475441.2005.9684215

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

Jannedy, S., Weirich, M., & Brunner, J. (2011). The effect of inferences on the perceptual categorization of Berlin German fricatives. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 962–965). Hong Kong: City University of Hong Kong.

Johnson, F. L., & Buttny, R. (1982). White listeners' responses to "sounding black" and "sounding white": The effects of message content on judgments about language. *Communications Monographs*, *49*, 33–49. https://doi.org/10.1080/03637758209376069

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech* *processing* (pp. 145–165). San Diego, CA: Academic. https://doi.org/10.1002/9780470757024.ch15

Jurafsky, D. (1993). Universals in the semantics of the diminutive. In J. S. Guenter, B. A. Kaiser, & C. C. Zoll (Eds.), *Proceedings of the Nineteenth Annual Meeting of the Berkeley Linguistics Society: General session and parasession on semantic typology and semantic universals* (pp. 423–426). Berkeley, CA: Berkeley Linguistics Society. https://doi.org/10.3765/bls.v19i1.1531

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Labov, W. (1972). *Language in the inner city*. Philadelphia: University of Pennsylvania Press.

Labov, W. (2001). *Principles of linguistic change: Vol. 2. Social factors*. Oxford, UK: Blackwell.

Ladegaard, H. J., & Bleses, D. (2003). Gender differences in young children's speech: The acquisition of sociolinguistic competence. *International Journal of Applied Linguistics*, *13*, 222–233. https://doi.org/10.1111/1473-4192.00045

Langstrof, C. (2014). Sociophonetic learning in L1 and L2. Unpublished habilitation thesis, University of Freiburg, Freiburg, Germany.

Lawson, E., Scobbie, J., & Stuart-Smith, J. (2011). The social stratification of tongue shape for post-vocalic/r/in Scottish English. *Journal of Sociolinguistics*, *1*, 256–268. https://doi.org/10.1111/j.1467-9841.2011.00464.x

Leung, J. H., & Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning*, *62*, 634–662. https://doi.org/10.1111/j.1467-9922.2011.00637.x

Leung, J. H., & Williams, J. N. (2013). Prior linguistic knowledge influences implicit language learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 2866–2871. Retrieved from https://escholarship.org/uc/item/6kh2b1dw

Lewicki, P., Hill, T., & Czyzewska, M. (1992). Nonconscious acquisition of information. *American Psychologist*, *47*, 796–801. https://doi.org/10.1037/0003-066x.47.6.796

MacFarlane, A. E., & Stuart-Smith, J. (2012). 'One of them sounds sort of Glasgow Uni-ish'. Social judgements and fine phonetic variation in Glasgow. *Lingua*, *122*, 764–778. https://doi.org/10.1016/j.lingua.2012.01.007

MathWorks, Inc. (2016). *MATLAB* (Version 9.0 (r2016a)) [Computer software]. Natick, MA: MathWorks Inc.

Maye, J., & Weiss, D. (2003). Statistical cues facilitate infants' discrimination of difficult phonetic contrasts. In B. Beachley, A. Brown, & F. Conlin (Eds.), *Proceedings of the 27th Annual Boston University Conference on Language Development* (Vol. 2, pp. 508–518). Somerville, MA: Cascadilla.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111. https://doi.org/10.1016/s0010-0277(01)00157-3

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*, 321–391. https://doi.org/10.1111/lang.12286
Mendoza-Denton, N. (1996). 'Muy macha': Gender and ideology in gang-girls' discourse about makeup. *Ethnos*, *61*, 47–63. https://doi.org/10.1080/00141844.1996.9981527
Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. https://doi.org/10.1146/annurev-psych-010416-044022
Milroy, J., & Milroy, L. (1993). Mechanisms of change in urban dialects: The role of class, social network and gender. *International Journal of Applied Linguistics*, *3*, 57–77. https://doi.org/10.1111/j.1473-4192.1993.tb00043.x
Milroy, L. (1980). *Language and social networks*. Oxford, UK: Blackwell.
Molnar, M., Ibáñez-Molina, A., & Carreiras, M. (2015). Interlocutor identity affects language activation in bilinguals. *Journal of Memory and Language*, *81*, 91–104. https://doi.org/10.1037/e524912015-064
Needle, J. M., & Pierrehumbert, J. B. (2018). Gendered associations of English morphology. *Laboratory Phonology*, *9*, https://doi.org/10.5334/labphon.134
Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*, 62–85. https://doi.org/10.1177/0261927×99018001005
Pharao, N., Maegaard, M., Møller, J. S., & Kristiansen, T. (2014). Indexical meanings of [s+] among Copenhagen youth: Social perception of a phonetic variant in different prosodic contexts. *Language in Society*, *4*, 1–31. https://doi.org/10.1017/s0047404513000857
Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, *4*, 516–530. https://doi.org/10.1016/j.wocn.2006.06.003
Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, *2*, 33–52. https://doi.org/10.1146/annurev-linguist-030514-125050
Pierrehumbert, J. B., Beckman, M. E., & Ladd, D. R. (2000). Conceptual foundations of phonology as a laboratory science. In N. Burton-Roberts, P. Carr, & G. Docherty (Eds.), *Phonological knowledge: Conceptual and empirical issues* (pp. 273–304). Oxford, UK: Oxford University Press.
Preston, D. R. (1996). Whaddayaknow?: The modes of folk linguistic awareness. *Language Awareness*, *5*, 40–74. https://doi.org/10.1080/09658416.1996.9959890
Qian, T., Jaeger, T., & Aslin, R. N. (2014). Implicit learning in a non-stationary environment: Knowing when it's Groundhog Day. Unpublished manuscript, University of Rochester, Rochester, NY.
R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rácz, P., Hay, J., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology*, *8*, 51. https://doi.org/10.3389/fpsyg.2017.00051

Roberts, G. (2017). The linguist's drosophila: Experiments in language change. *Linguistics Vanguard*, *3*, https://doi.org/10.1515/lingvan-2016-0086

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114. https://doi.org/10.1016/j.cogpsych.2017.02.004

Sankoff, G., & Blondeau, H. (2007). Language change across the lifespan:/r/in Montreal French. *Language*, *83*, 560–588. Retrieved from http://doi.org/10.1353/lan.2007.0106

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*, 411–417. https://doi.org/10.1016/j.tics.2010.06.006

Silverstein, M. (2009). The limits of awareness. In A. Duranti (Ed.), *Linguistic anthropology: A reader* (Vol. 1, pp. 382–401). Oxford, UK: Blackwell.

Smith, J., Durham, M., & Richards, H. (2013). The social and linguistic in the acquisition of sociolinguistic norms: Caregivers, children, and variation. *Linguistics*, *51*, 285–324. https://doi.org/10.1515/ling-2013-0012

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

Timmins, C., Tweedie, F., & Stuart-Smith, J. (2004). *Accent change in Glaswegian (1997 corpus): Results for consonant variables*. Glasgow, UK: Department of English Language, University of Glasgow. Retrieved from https://www.academia.edu/9432814/Accent_change_in_Glaswegian_1997_corpus_Results_for_Consonant_Variables

Trechter, S. (1995). Categorical gender myths in Native America: Gender deictics in Lakhota. *Issues in Applied Linguistics*, *6*(1), 5–22. Retrieved from https://escholarship.org/uc/item/2257330q

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge, UK: Cambridge University Press.

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. https://doi.org/10.1162/jocn.2008.20054

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

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* [Computer software]. New York, NY: Springer-Verlag. Retrieved from http://ggplot2.tidyverseorg

Woodward, A. L., & Markman, E. M. (1998). Early word learning. In W. Damon (Ed.), *Handbook of child psychology (Vol. 2): Cognition, perception, and language*. Hoboken, NJ: Wiley.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Model Fitting.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Some Language-Context Associations Are Harder to Learn Than Others: An Artificial Language Study

What This Research Was About and Why It Is Important

We associate words with the people who usually say them. (For example, we intuitively expect that "confectionery" will be used by older speakers, while "selfie" will be used by younger ones.) This is known as "higher indexical knowledge of language" and it can be very complex. Members of a speech community (such as a village, a high school, or a bowling team) use it to perform their social selves and to navigate their social world. Scientists know relatively little about the way we come by this knowledge. Can we learn any and all such associations? Are we attuned to socially relevant ones? Given the problem's complexity, these questions are difficult to study in the wild. The authors of this study opted for a laboratory experiment, an artificial language task that trains participants to pay attention to certain associations between language and context and ignore others. They found that not all language-context associations in their task were equally easy to learn. Participants were relatively good at picking up that a word pattern goes with the speaker's gender. They found it harder to learn that a word pattern goes with whether the speaker is an adult or a child, for instance. This shows that indexical learning is *selective*. We are more attuned to some contexts than others.

What the Researchers Did

The study used a simple artificial language learning task, run on an online platform (Amazon Mechanical Turk).

- 474 American English-speaking adult participants had to learn variations of the same artificial language.
- In all such variations, word use was associated with virtual conversation partners. Some partners used one word, others used a different word to describe the same things. This echoes how such associations play out in real life.
- In some cases, the key to this association was something socially relevant: the partner's gender, age, or ethnicity. In other cases, it was something socially irrelevant: which direction the partner is facing.

What the Researchers Found

- Participants found it relatively easy to learn that women and men use different words. They found it harder to learn that adults and children do so. They were mostly unsuccessful in learning that people facing different directions would use different words as well.
- In cases where learning the association was relatively easy, participants could also extend (generalize) the artificial language: for instance, they assumed that (new) women would use the same word pattern for new words.
- If the association was harder, participants struggled to extend it: they could learn that adults and children use different word patterns but could not extend this successfully to words they had not seen before.

Things to Consider

- An artificial language learning task does not approximate the rich complexity of real-life interactions. However, its abstract nature allows us to study specific aspects of higher indexical knowledge and how individuals learn it.
- The results suggest that our prior beliefs about social life contribute directly to how we recognize associations between language and context: we pay more attention to socially relevant dimensions and less to irrelevant ones.
- While adults are very good at keeping track of details when creating categories in general, it seems that, in learning indexical language use, they heavily rely on their past experience of what is socially relevant.
- The results also shed light on the mechanisms of learning a new association. It seems that it takes a while to completely generalize a relationship between some contexts and some pieces of language to all relevant contexts.

- Results from adults are interesting because they show learning higher indexical knowledge continues throughout our lifetime. However, an important follow-up would be to study how young children, on the cusp of grasping the importance of social relations, would behave to explore the learning of higher indexical knowledge across the lifespan.

Materials and data: Materials are available at https://github.com/nzilbb/roofRunner and data are available at https://doi.org/10.5281/zenodo.3519395

How to cite this summary: Rácz, P., Hay, J., Pierrehumbert, J. B. (2019). Some language-context associations are harder to learn than others: An artificial language study. *OASIS summary* of Rácz et al. in *Language Learning*. https://oasis-database.org