Not all indexical cues are equal: Differential sensitivity to dimensions of indexical meaning in an artificial language

Péter Rácz
Cognitive Developmental Center,
Central European University

Jennifer B. Hay
New Zealand Institute of Language Brain and Behaviour,
University of Canterbury

Janet B. Pierrehumbert
Department of Engineering Science,
University of Oxford

December 13, 2019

Draft version, to appear in Language Learning, 2020

Abstract

This paper investigates the learning of indexical features by English-speaking adults using a novel experimental paradigm. Participants learn an allomorphy pattern cued by a given social context. The social contexts are represented by conversation partners who differ by age, ethnicity and/or gender and are positioned in various ways. The results show that, after training, participants are able to learn that different types of conversation partners prefer different types of allomorphs, but that learning and generalisation hinges on the social relevance of the cue represented by the conversation partner.
These results suggest that the relevance of cues in the individual’s past social experience influences their storage and learnability even at very early stages of learning a word pattern.

Keywords: sociolinguistics, exemplar theory, morphology, gamification, learning

1 Introduction

Many words carry statistical associations with non-linguistic contexts – they might be used more by some types of speakers than others, or be preferred with certain interlocutors, or in certain contexts. In New Zealand English, for example, the word ‘lovely’ is statistically more likely to be produced by a female speaker than a male speaker, and the word ‘confectionery’ is unlikely to be produced by a speaker who is young (Hay et al. 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.) Non-random associations between linguistic and non-linguistic contexts can play a crucial role in early word learning (Woodward and Markman, 1998) and continue to influence language processing throughout the lifespan (Chater and 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 and Itti 2013). On the other hand, we know a lot about the ways in which the context is indexed in linguistic conventions and how this contributes to language variation and change (see e.g. Niedzielski 1999; Hay and Drager 2010; Bucholtz 1999).

The missing piece here is how people learn these associations between language and the non-linguistic context – the subject of this paper. We use an artificial language learning task to investigate how indexical associations in language are learned. The great benefit of such a paradigm is that it allows us to operationalise 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 et al., 2014; Roberts, 2017). The approach necessarily entails a number of abstractions and cannot capture
the richness of real-world indexical associations. This only shows that a multi-faceted problem like indexicality needs to be approached from multiple angles. The main contribution of the work reported here is that it adds to the toolkit for studying the development of indexical associations.

In this paper, we use a novel experimental paradigm and artificial stimuli to study how adults learn and process social-contextual cues to linguistic variation. We look at the relationship between a non-linguistic context and a linguistic pattern in a simple learning task. In the task, adult speakers learn to associate two specific morpho-phonological patterns with two contexts. This work follows up on Rácz et al. (2017). 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; here we go on to focus on different types of social contexts and the precise mechanics of learning associations with these contexts. Our aim is to test whether certain types of context-language associations are learned faster than others, and whether these are also generalised more easily to new linguistic and non-linguistic contexts.

Our task tests these questions by teaching an association between language and context. We first discuss existing work in the two areas outlined above: on the role of the context in general category learning and extant context-language associations in the ‘wild’. We also briefly survey past work on learning indexical associations in the lab.

2 Background

2.1 Contextual learning

People are able to rely on the context in learning tasks. They can create associations between linguistic or non-linguistic 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, there is plenty of evidence 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 are learned underwater are more accurately recalled underwater. Hay et al. (2017), in a related study, showed location-specific effects on speech perception in the laboratory and in a car. A non-linguistic example of contextual
learning is reported by Qian et al. (2014), who showed that in a ‘whack-a-mole’ type game, players are faster at predicting the location of the mole if the location is probabilistically cued by moving background images that the player is not overtly oriented to. Much earlier work on implicit learning was reviewed by Lewicki et al. (1992). 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 shows that learned associations between context and language usage play an important role. Van Berkum et al. (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 (like ‘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 (like ‘The Earth revolves around the trouble’). This suggests very early involvement of contextual information in sentence processing. Furthermore, within the set of incongruous sentences, they find additional differences in neural activity for sentences which are incongruous with a female or male speaker (such as ‘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.

Another example of the context feeding into language processing comes from Molnar et al. (2015). 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 speak for 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 speculate that learner experience is vital here. One possible explanation for their results, they suggest, is that
the critical factor driving implicit learning of form-meaning connections is not their availability in itself. Rather, it is their availability to grammatical processes and representations, based on the prior linguistic knowledge of the individual. 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.

2.2 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 only provide 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 factors influence linguistic variation at all levels – from phonetic to syntactic variation. We rely on the social meaning of lexical items both in speech perception (Giles et al., 1973; Niedzielski, 1999; Hay et al., 2006; Foulkes et al., 2010; Jannedy et al., 2011; Campbell-Kibler, 2011; MacFarlane and Stuart-Smith, 2012; Pharao et al., 2014) and speech production (Labov, 1972; Trudgill, 1974; Milroy, 1980; Eckert, 2000; Labov, 2001; Timmins et al., 2004; Foulkes and Docherty, 2006; Hay and Drager, 2007; Lawson et al., 2011). Speakers are able to keep track of the effect of context even at the word level (Pierrehumbert, 2016; Pierrehumbert et al., 2000; Hay et al., 2019).

Speakers of different dialects will focus on and learn different linguistic details (Cohn et al., 1999), while speaker awareness of contextual information on all levels can be very imprecise, and usually 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 relatively low.

The significance of social contexts varies in sociolinguistic variation, echoing work on contextual learning. Some factors, like gender (Cheshire, 2002; Milroy and Milroy, 1993), age (Sankoff and Blondeau, 2007; Walker and Hay,
2011), and ethnicity (see e.g. Johnson and 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. Mendoza-Denton 1996; Habick 1991).

2.3 Sociolinguistic learning

Of course, despite our affinity for learning context-language associations we do not start with a perfect knowledge of sociolinguistic variation right away – it is learned along with the other, denotative and structural aspects of language.

Learning the associations present between non-linguistic contexts and linguistic patterns starts early (Foulkes et al., 2010; Smith et al., 2013) and continues into adulthood, as evidenced by ongoing changes in the linguistic variation of the individual, mirroring changes in the community (Harrington et al., 2000). The mechanisms through which knowledge about social variation is acquired are not fully understood, but 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 and Foulkes, 2013; Pierrehumbert, 2006). Indeed, modern theories of the mental lexicon, the storage of linguistic forms, tend to argue that non-linguistic and linguistic contextual information plays 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 et al., 2007).

Docherty et al. (2013) investigated the learnability of several types of sociophonetic associations using a methodological paradigm which involves passive exposure to words produced by two ‘tribes’. Across different experiments, tribes differed in terms of the phonetic markers that distinguish them. In a subsequent test phase, participants listened to the same recordings they had been exposed to and were asked to overtly label them as originating from ‘tribe 1’ and ‘tribe 2’ which they did with a degree of success. A number of further experiments using this paradigm were reported by Langstrof (2014). These works focus on existing types of sociophonetic variation, and restrict tests of pattern learning to words already encountered in training.

Docherty, Langstrof, and Foulkes’ most relevant result to the present work is the demonstration that adult listeners do form associations between lin-
guistic 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. While these works show that sociolinguistic learning is possible in the laboratory, the types of non-linguistic contexts remain deliberately artificial. This is despite the existing assumption that some non-linguistic contexts are easier to recognise and learn than others. For instance, Foulkes (2010) hypothesised that some types of social-contextual properties should be more readily transmitted and learnable than others, due to the variable frequency with which properties have been relevant in individuals’ past experience. This ties in with the observation that some non-linguistic contexts are more strongly associated with sociolinguistic variation than others. Foulkes goes on to identify speaker sex as one of the very earliest learned socio-indexical associations. Of course, speaker sex is marked in the grammar in Indo-European languages.

Existing work provides evidence that a linguistic association with gender is indeed learnable in the laboratory. Samara et al. (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’, is 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 generalise these associations to complex pseudowords. Hay et al. (2019) show that a word’s associations with both gender and age can affect lexical access patterns.

It is implicit to much of the work on sociolinguistic learning that associations between language and a non-linguistic context are generalised. Such associations are, for the most part, formed between context and linguistic categories, and not restricted to individual items. The notion of the 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 pointed out by Pierrehumbert (2006), 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. There is little in the literature that directly probes the mechanics of generalization of learned socio-indexical associations. Sneller and Roberts (2018), for instance, showed that the adoption of new sociolinguistic variants in the laboratory hinges on their contextual associations, hinting at the joint role of context and generalisation in sociolinguistic learning, but they do not address the specifics of within-context generalisation directly.

In Rácz et al. (2017) we looked at the process of learning social meaning by comparing types of non-linguistic contexts. The focus of Rácz et al. (2017) was the process of learning, contrasting learning in a strictly linguistic versus a mixed context, exploring the influence of the linguistic context on learning an association with a non-linguistic context, and the link between training and generalisation strength. The results showed that learning and generalisation are not robustly different in a linguistic and in a non-linguistic context. At the same time, instance-based generalisations were more important for a linguistic context, whereas non-linguistic associations were treated more generally. While it is true for all the contexts discussed in Rácz et al. (2017) that there were participants who are able to rely on them, the share of successful learners clearly hinged on the type of the context. In particular, participants were best at associations with a gender-based difference between conversation partners. The varying strength of non-linguistic contexts is further explored in this paper. The focus of this paper is variation in learning accuracy and the extent of generalisation across various non-linguistic contexts.

3 Rationale of the Current Study

Previous research has demonstrated that adults are able to keep track of a large amount of very detailed non-linguistic context in language processing (Needle and 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 and Williams, 2012). These aspects of contextual learning are reflected in sociolinguistic variation: it utilises a wide array of contexts, some of which are more readily associated with linguistic variation than others. We have seen that sociolinguistic learning is a lifelong process, one that can be studied in the laboratory. Several works, including our own, have shown that participants can learn a
gender-based contextual distinction in an experimental setting. These results bring new questions to the fore. Are gendered patterns learnable in the laboratory because gender is highly salient, as evidenced by its early acquisition (Ladegaard and Bleses, 2003), as well as behavioural and neural evidence (Cheshire, 2002; Van Berkum et al., 2008)? As we discuss in the Background section, speaker age and ethnicity are also prevalent factors in sociolinguistic variation. Is this because these non-linguistic 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 generalise?

This study sets out to test these questions couched in the broader frameworks of sociolinguistic variation and sociolinguistic learning. We hope to further demonstrate that sociolinguistic learning can not only be attested but also dissected in the laboratory, and that an artificial language task, despite the necessary simplifications, can add to the toolkit of studying language variation and change by shedding further light on its mechanics.

This leads us to our starting hypotheses:

1. A non-linguistic contextual cue that is socially salient (like conversation partner gender) is easier to learn than an irrelevant cue (like conversation partner spatial orientation). (Our 2017 paper found 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 generalisation: a pattern that is learned with a socially salient contextual cue is easier to generalise to new linguistic and non-linguistic contexts.

4 Methods
4.1 Design

The backbone of our design is a simple artificial language learning task, adapted from (Rácz et al., 2017). It consists of a training phase followed by a test phase. In training, the player takes up the role of a chicken that flies from roof to roof, going towards a destination (its nest). The chicken meets
 various conversation partners, one in each trial. The conversation partner shows the chicken a prompt image with an accompanying name. The chicken has to respond to this prompt with a related named image. The response image is given, but the player has to pick the name out of two options. A correct answer means the chicken can go on, an incorrect answer results in going back two trials. The player has to pick the correct answer to every question to finish training.

The upper image in Figure 1 shows the layout of a training trial. The trials are entirely visual. On the left, we see the chicken. On the right, we see the conversation partner, an adult female wearing a yellow top. The prompt is tied to the conversation partner by a speech bubble: it 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, while 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, as no clue is given to the correct answer. For the sake of this example, let the correct answer be ‘fenpel’. Picking this 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: it is now an adult male in a black t-shirt.

Choosing the wrong answer (rikpel, with the -pel suffix) will send the player back two trials. Choosing the correct answer (‘rikfis’) will take the player to the next trial (Figure 1, third image). Here, we see a new named object 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 test phase follows the training phase. A test trial is similar to a training trial except that it is night rather than day; the chicken is no longer present, and the player receives no feedback. The test phase has items and conversation partners familiar from the training phase as well as entirely novel ones, as seen in Figure 2. The aim of test is to see whether the pattern learned in the training phase carries over and whether it is generalised to items and conversation partners not seen in training; in our example above,
Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

Figure 1: The layout of the training phase of the artificial language learning task
a previously unseen item and a male partner not seen in the training phase should still lead the player to pick the ‘male’ suffix, ‘-fis’.

Our 2017 study employed a basic design that uses 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 is how these images were grouped. In the so-called gender condition outlined above, the correct answer (here: ‘-pel’ or ‘-fis’) depended on whether the conversation partner was a woman or a man. Which way they face was irrelevant. In the corresponding view condition, the correct answer depended on which way the conversation partner faces: people facing front will prefer ‘-pel’, while conversation partners facing sideways will prefer ‘-fis’, for instance. A player was randomly assigned either into the gender or the view condition.

In the 2017 study we found that the gender distinction is learned and generalised much more easily than the view distinction. This extended to new items and new conversation partners (from women to girls and men to boys, for instance). In the current study, we expand the set of conversation partners in order to test more groupings beyond gender and spatial orientation.

4.2 Materials

Players always saw four conversation partners in the training phase and four additional ones in the test phase. Who these partners were varied from player
Rácz, Hay & Pierrehumbert (*to appear*): Not all indexical cues are equal

Figure 3: The four conversation partners in the 2017 study: they differ in gender.

to player. Conversation partners were drawn from a complete set, seen in Figure 4. They could 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, four of a kind were new. The new partners introduce a new dimension of contrast – for example, if training was with adults, the new partners may all be children.

To reiterate the example from the previous section, the *main cue* was *gender* (this determines which answer is correct in training), the *competitor cue* was *view* (the partner’s spatial orientation; this should be ignored). The training phase introduced new partners, drawn from the pool of images unused in the training – for example if partners in the training phase were adults, the new, unfamiliar partners in the test phase could have been children.

So far we discussed conceptual similarities across our conversation partner images (such as same gender, age, etc.). However, these images are also perceptually similar to each other in shape and colour and this partially overlaps with conceptual similarities. We are exclusively interested in conceptual similarity and we thus used post-hoc checks to control for the effect
Figure 4: The eight conversation partners in the current study: they differ in gender, age, ethnicity, and view (spatial orientation).
Participants encountered six items in the training phase and these six items and six additional items in the test phase. Our twelve item images were distributed randomly across the test phase and the training phase for each participant. Suffixes were picked and stems were assigned to images randomly for each participant.

Item names in the task were built from an artificial language. Since the focus of the task is social association, the artificial language itself was deliberately simple. For each participant, syllables were randomly drawn from a finite syllabary (see Table 1). Two syllables became suffixes, and the rest were randomly assigned as names to the two times six items.

The syllabary reflected the following design principles: (i) the syllables should be distinctive; (ii) they should consist of a small set of frequent letters; (iii) they should be easy to pronounce for our participants, who are American English speakers; (iv) the consonant clusters in the two-syllable words should cue English word boundaries in a uniform manner. Our aim was to provide a relatively optimal set that balances these considerations.

We used the diminutive as the contextually cued morphological category for most of our participants. This is because the diminutive is a common, iconic pattern that is easy to interpret visually. However, it is highly variable in English and it has strong associations with gender in many languages (Jurafsky, 2012). In order to make our findings more robust, we replicated two conditions using plural instead of diminutive as the morphological category. In these replicated conditions, participants performed the exact same task except that diminutive images were replaced with plural ones. The words were similar, the implied meaning different.

In the diminutive condition, the representation of the target item is 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 is a picture of three of the target items, normally scaled, instead of one diminutive version.
4.3 Participants

The experiment was hosted on Amazon Mechanical Turk. The platform has been used successfully (albeit with caveats) in behavioural research (Crump et al., 2013) and provides a participant pool that is more representative than the typical convenience sample (Berinsky et al., 2012).

Participants were run in three large batches, each several weeks apart in 2014 (see Rácz et al. 2017) and 2015. Participants were paid 3 dollars upon completion of the task. Participant IP addresses were restricted to the United States and participants had to be native speakers of English. Participant information was collected using a pre-task questionnaire.

474 participants took part in the experiment. 11 people were removed 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 took 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, with a standard deviation of 2.4 minutes. The fastest participant finished in 1.6 minutes, the slowest (after filtering) in 18.93 minutes.

The duration of individual trials, however, provides an unreliable metric since participants play the game on their own computers and not in a laboratory setting. A participant who gets distracted or stands up to make a cup of tea will take longer to finish a trial, much the same way as a participant who has a 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 has a minimum of 24 (6 items seen with 4 conversation partners, and the participant responds correctly to every combination) and a long tail.

Participant sample sizes vary across the conditions. This was partly due to variability in exclusion rates and partly to shifting experimental protocols—the minimum sample size of 30 per condition. We took additional steps to make sure that this does not affect the results, discussed in the Results section.

Since we do assume that the training phase is longer in certain across-participant conditions than others, and since we wanted outlier thresholds to reflect this, we took every condition separately and then removed the slowest 2.5% based on training trial count threshold (within conditions)—this
leaves 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 were this slow.

### 4.4 Procedure

The training phase consisted of six images with four conversation partners (24 trials), whereas the test phase consisted of twelve images (six unfamiliar and six familiar) with eight conversation partners (four familiar and four unfamiliar) (48 trials). Each participant saw 72 unique trials in total.

The task had three across-participant factors, 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 factors, specific to the test phase: whether the conversation partner or the target item was *familiar* – that is, whether it had been encountered in training.

Table 2 shows the existing main cue / competitor cue combinations in the experiment – these are the effective across-participant conditions.

| Competitor cue | age | ethnicity | gender | view |
|----------------|-----|-----------|--------|------|
| Main cue       | age | ✓         | ✓      | ✓    |
|                | ethnicity | ✓         | ✓*     |       |
|                | gender | ✓         | ✓*     |     |
|                | view    | ✓         |        | ✓    |

Table 2: Figure 4 is helpful in interpreting Table 2. In the table, a check mark means that the combination has been tested in the experiment. Some combinations are not possible (a cue cannot be a main cue and a competitor cue at the same time) while 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 one was used to introduce new conversation partners in test. This was never the view cue to enforce the interpretation that new conversation partners are different individuals.

Table 3 shows the participant counts and summary demographics across the across-participant conditions.
### Table 3: Summary statistics

| main cue (competitor cue)   | pattern | participant count | men ratio | mean age | sd age  |
|-----------------------------|---------|-------------------|-----------|----------|---------|
| age (ethnicity)             | dim     | 30                | 0.53      | 30.59    | 9.97    |
| age (gender)                | dim     | 36                | 0.34      | 30.77    | 6.88    |
| age (view)                  | dim     | 49                | 0.37      | 34.55    | 11.58   |
| ethnicity (age)             | dim     | 30                | 0.38      | 33.28    | 13.40   |
| ethnicity (gender)          | dim     | 26                | 0.64      | 32.40    | 12.50   |
| ethnicity (gender)          | plural  | 35                | 0.54      | 30.54    | 8.46    |
| gender (age)                | dim     | 37                | 0.59      | 33.00    | 10.28   |
| gender (ethnicity)          | dim     | 26                | 0.54      | 29.54    | 8.38    |
| gender (ethnicity)          | plural  | 31                | 0.39      | 32.45    | 7.83    |
| gender (view)               | dim     | 47                | 0.38      | 33.74    | 11.58   |
| view (age)                  | dim     | 40                | 0.40      | 32.90    | 9.54    |
| view (gender)               | dim     | 48                | 0.48      | 30.47    | 9.52    |

#### 4.5 Data analysis

We report results primarily from the test phase. (Participants’ training performance is used as a predictor of their accuracy in test.) 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 results.

We used the `lme4` package implemented in R (Bates et al., 2015; R Core Team, 2018) for model fitting and `ggplot` (Wickham, 2016) for plots. Results for the gender (view) and view (gender) conditions were also reported in Rácz et al. (2017). Model 1 is fit on test data for the diminutive. Model 2 compares diminutive and plural data. Model 3 tests the effect of perceptual versus conceptual distance between conversation partner images in the test phase. Model 4 refits Model 1 while re-sampling 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. We tested the robustness of individual main terms the same way.

A number of criteria exist for model comparison. We relied on chi-square goodness-of-fit tests to select the best model. The model fitting process is outlined in detail in the Appendix. We report the model with all main terms and robust interactions. We include the main terms since they express aspects of the experimental design. The interactions are more exploratory and post hoc, and so we only include robust ones in the reported model.
For Model 1, Model 2, and Model 3, we report the results of goodness-of-fit tests and provide more details in the Appendix. For Model 4, we ran a Monte Carlo simulation and provide results with an error threshold.

### 4.6 Results

Our hypotheses were that participant accuracy in test varies across (1) main cues and competitor cues and (2) for previously seen items / conversation partners compared to those only present in test. In our 2017 paper, we investigate two contextual cues, meaning that the main cue always implies the competitor cue. In the current study main cue and competitor cue are independent and considered separately.

Model 1 was fit to test these hypotheses for the diminutive pattern using multilevel binomial generalised linear regression. The outcome was whether the participant picked the correct name for the diminutive item in the test phase. 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 participant’s training trial count as a predictor. The model also had a participant grouping factor. Since item images and names are randomised, we did not include an item grouping factor.

Figure 5 shows the estimates for the best model for the results of the test data (Model 1.2) with Wald 95% confidence intervals. (For the values please consult the Appendix.) The Wald 95% confidence intervals capture the certainty of the estimates: where the interval excludes zero, we can be 95% certain that the true difference is non-zero.

We first walk the reader through the term estimates, then turn to visualisations of our data to expound on the relevant patterns, then provide a summary of how these patterns relate to our hypotheses.

What we see here is that accurate response to a test trial varies across main cue type, and that this is mediated by familiarity with the conversation partner ($\chi^2 = 155.4, p < 0.001$). Participants who finished training in more trials are also less accurate in test ($\chi^2 = 141.94, p < 0.001$). The competitor cue ($\chi^2 = 4.08, p = 0.25$) and familiarity with items ($\chi^2 = 0.81, p = 0.37$) make no robust difference in determining participant accuracy.

Main cue is an unordered factor and the implications of term estimates are relatively hard to interpret. We turn to visualisations of the test data to
Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

Figure 5: Estimates of model 1 on diminutive test data

shed more light on participant behaviour.

The main effect of cue type can be seen in Figure 6. The figure shows participant average accuracy in the test phase across main cue type (upper panel) or competitor cue type (lower panel). The individual coloured points are participant averages, the violins show their distributions across cue types, and the black points show the means of these distributions.

The first interesting aspect of participant distributions is that they vary across main cue type. Looking at main cues, participants who have to learn the view cue in the training phase are the least accurate, overall, in test. Participants learning the gender cue are the most accurate, with age and ethnicity in between. Looking at competitor cues, participants who have to ignore the view cue are the most accurate, while participants who have to ignore the gender cue are the least accurate. However, variation across competitor cues is 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 do not see 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 is clustered
Figure 6: Participant averages across main cue (upper) or competitor cue type (lower) in test. Coloured dots are jittered participant averages, black dots are means for cue types.
Figure 7: Participant averages across main cue and familiarity with the conversation partner in test. The figure breaks down the upper panel of Figure 6. Each violin is split into two: one for unfamiliar and one for familiar conversation partners in test. Each grey line represents one participant.
round 0.5 (chance level), while another group is close to 1 (ceiling). It is likely that participants who understood the rule they have to learn in the training phase are the ones that are very accurate in test, while those that kept guessing and passed the training phase by rote learning the correct answers are the ones who are also guessing in test. It is, in fact, the proportion of these two sets across main cue type that shifts overall means: more participants figure out the gender rule than the view rule. This implies that participant success in test correlates with participant success in training. Indeed, the training trial count is a significant main effect in the model.

Model 1 indicates that participant accuracy across main cues is mediated by familiarity with the conversation partner, as illustrated in Figure 7. This figure replicates the upper panel of Figure 6, split according to whether the conversation partner is 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. What we see is that, for view and gender, familiarity with the conversation partner makes no difference in participant accuracy. Participant averages move up and down in the split to some extent, but no clear pattern is visible. In contrast, for both age and especially ethnicity, some participants are drastically more accurate in test trials with conversation partners that are familiar from training. This is despite the fact that new conversation partners share the same grouping characteristics (for instance, they are also children or adults).

Participants who have to learn a contextual distinction that is not supported by prior knowledge (the view cue) will mostly keep guessing in the test phase overall, hence no improvement with familiar conversation partners. For participants who manage to learn the gender distinction in training, generalisation is already so complete that they are at ceiling accuracy with both old and new conversation partners.

The two categories in between are eminently learnable (at least for a considerable minority of participants) but generalisation to new conversation partners is not straightforward – participants benefit from familiarity. This challenges the interpretation we allude to above, namely that participants either learn 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 is successful at learning meaningful contextual associations (gender, age, ethnicity) than a non-meaningful association (view). Gender, in particular, is the easiest to learn out of the three. Gender-based associations are generalised straightforwardly
to new conversation partners, while those based on *age* and *ethnicity* are also generalised, but to a lesser degree. This indicates that a robustly learnable distinction is also robust to generalise.

In order to assess the robustness of these results, we fit a number of secondary models to test potential confounding factors. First, we examined the experimental conditions for which we collected both diminutive and plural data. Model 2 was fit on cue types tested with both the diminutive and the plural patterns: *ethnicity / gender* (here, *gender* is either main cue, with *ethnicity* as the competitor, or the other way round). We found that pattern type does not explain more variation in the data ($\chi^2 = 0.01, p = 0.94$).

Second, we verified that the results are 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 has affected participant behaviour. Though these two are intertwined, we want participants to react to conceptual similarities (e.g. ‘same gender’) without the interference of perceptual similarities (‘same height’). We used a signal processing metric (Levenshtein distance between the images) as a measure of ‘visual differences’. This is because reliance on human raters would necessarily invoke the social-conceptual distances as well in determining visual distance.

We calculated the Levenshtein distance for all conversation partner image pairs in our training data. We used the Image Processing Toolbox of Matlab (MATLAB, 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 that the participant is trained on versus the category that the participant needs to ignore. If perceptual distance is relevant, participants should have higher accuracy in learning associations with more contrastive conversation partner categories.

Model 3, a multilevel binomial generalised linear regression model, was fit 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 factors.

The model was compared to the best fit of Model 1 – excluding training trial count – to see which one explains more variation in the test data: the model that relies on perceptual distance or the one that relies on conceptual distance. What we found is that the model which relies on conceptual dis-
tance gives a much better fit than the one that relies on perceptual distance ($\chi^2 = 173.55$, $p < 0.001$).

Finally, we verified that the uneven sample size was not responsible for our key result. Modelling attempt 4 is a replication of Model 1 with resampled participant sets. Sample size varies across conditions. In order to make sure that our results are 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, \textit{ethnicity (gender)}: 26. We fit a multilevel model with the interaction of \textit{main cue} and \textit{familiarity with the conversation partner} on each resampled data set. Using a $z$ value of 1.8 as a cut-off, we find that the crucial interaction remains robust with \textit{age} in 38/100 models and with \textit{ethnicity} in 96/100 models. The interaction of \textit{view} or \textit{gender} with conversation partner is robust in 0/100 models. This indicates that the interaction specified in the best fit of Model 1 is not an artefact of the variance in sample sizes.

In summary, players struggle to learn an association with the view cue. They are likely to easily learn, and completely generalize, an association with gender. Associations with age and with ethnicity are learned moderately well, and in our data they show evidence of some instance-specific learning, with the process of generalization still underway.

5 Discussion

The task of learning an association between a non-linguistic context and an allomorphy pattern is a hard but not impossible one for our participants. Participant accuracy reflects real-life sociolinguistic knowledge brought into the experiment. Training takes longer and test accuracy is lower with a less socially relevant cue as compared to socially relevant ones. Within our small set of socially relevant contextual cues, participant behaviour in the test phase indicates that the context-pattern association is initially rote-learned and then generalised to new contexts (new conversation partners). This generalisation is easier with a socially more robust cue.

In effect, we can think of training with any of the socially relevant cues (\textit{gender}, \textit{age}, and \textit{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 test, the participant has to recognise that the new individuals they come across share characteristics with the old ones seen in training:
if the two children prefer suffix A in training, all children will prefer it in test, too. For our participants, this particular generalisation is harder than recognising that if the two female characters prefer suffix A in training, all female characters will prefer it in test.

It is interesting that the competitor cue did not robustly affect participant accuracy. We speculate that this was because the paradigm only reinforced the main cue, which could have an unbalancing effect on cue competition by focussing attention. Alternatively, it is possible that participants who focussed on the competitor cue ended up making many mistakes in training and getting feedback that is not informative to them. 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 demonstrates the use of prior non-linguistic knowledge of which social differences are commonly signalled by linguistic differences, matching work discussed in the Background section. Spatial orientation, which could play a role in resolving deixis, but has no social salience, is the hardest to associate with a linguistic pattern in this task. This remains true despite the design’s apparent simplicity, relying on a small artificial language and exaggerated cartoon representations of extant social constructions. It shows how adult sociolinguistic learning can be dissected using an artificial language task: It can be seen as a process that starts as an association of a linguistic pattern with a specific non-linguistic context, gradually generalised to other, similar contexts. Of course, how this generalisation unfolds in real life and its possible limits are beyond the scope of this study.

The task design makes it clear that allomorph selection is a response to the conversation partner, and that incorrect responses impede success. While the primary aim was to render the task as straightforward as possible, this setup also has real-life precedents: for example, in French, incorrect marking of the gender of an adjective can lead to incomprehension (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 is entirely visual and responses are in a forced-choice format. An open format would have likely resulted in a different pattern of responses, but it would have also required the participants to effectively memorise the entire syllabary, rendering an already challenging task even more difficult.

With this in mind, we can 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 generalised (from familiar conversation partners to new ones). These results expand on our 2017 study which showed that it is possible to study contextual language learning using artificial language methods. We expand on this study by broadening the range of contexts that we investigate and find that their real-life prevalence manifests in how participants learn and generalise 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 factors also fare differently. We thus have 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 sex as one of the very earliest learned socio-indexical associations, and in our experimental paradigm, we have shown that – of the contextual variables we tested – gender is the most easily attended to and learned by adults. Needless to say, the 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 show that both adults and children can learn a gender-based association in an experimental setting, and Needle and Pierrehumbert (2018), who show that the gendered associations of suffixes for American English speakers carry over to pseudowords. In our task, both the stems and the suffixes are pseudowords, and the gendered association is established during the task. Learning, at least within the task, still takes 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 implicit socio-contextual learning. Conversation partner gender is very likely amongst the most frequent variables our participants have encountered in terms of conditioning linguistic variation. It is also marked explicitly in 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 factor which 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.

Certainly, these results accentuate the need for a more nuanced understanding of mechanisms of linguistic variation and transmission. Much of the literature documents how social factors are reflected and constructed through variation in speech, and much speculation and modelling relates to how these associations emerge and are transmitted (see Sneller and Roberts 2018; Foulkes and Hay 2015). A missing piece in this literature, however, is that not all social factors are equal in terms of how much we attend to them and how much we store them when processing and learning language. Salient social groupings are important not only in influencing the nuanced ways in which we produce and construct language, but are also differentially implicated in the very information that we store when we 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- and individual-specific social factors are at play (Eckert, 2000).

Our results also point to interactions between the robustness of a non-linguistic contextual cue and the learning process itself, as our participants find it easier to generalise a more robust context. The present study cannot account for the entire process of learning sociolinguistic variation, particularly because our participants are all over 18 and our design uses one-to-one correspondences between context and patterns (women always use ‘-ril’ etc.). Consistency of input has a huge effect on learning associations, and child and adult learners react to input variability very differently (Kam and Newport, 2009; Hudson Kam and Newport, 2005). While specific instances of e.g. linguistic address 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 multi-faceted, with playful and metalinguistic uses present (see e.g. Trechter 1995). In American English, the native language of our participants, gender is practically never associated categorically with social language use (Eckert and McConnell-Ginet, 1992).

In addition, the closely related results of Samara et al. (2017) show that both adults and children are able to generalise linguistic cues of speaker identity, even if these cues are probabilistically associated with the non-linguistic context. It is all the more remarkable that we see learning and generalisation vary considerably as a function of the type of social context in
our study, despite the absolute consistency of input.

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 not addressed in this paper (Maye et al. 2002; Maye and Weiss 2003, though see Atkinson et al. 2015). In our study, the number of nonce words and the number of non-linguistic context types is relatively low – differences in learning accuracy still emerge. (In Rácz et al. 2017, we show that training with 18 instead of 6 items improves participant accuracy in test.)

Our work uses an artificial language to investigate the pre-existence 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 classic work (1958) and artificial language paradigms have been valuable tools of investigating complex linguistic phenomena in a controlled environment (see Scott-Phillips and Kirby 2010; Roberts 2017).

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. What we hope to have shown is that a relatively simple paradigm can provide insights 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 very beginnings of understanding how intersecting social factors are implicated in, and impact socio-contextual learning.

Supplementary Material

All data and code available at 10.5281/zenodo.3519395.

Acknowledgements

The authors would like to thank Chun Liang Chan, Patrick LaShell, Clay Beckner, Vica Papp, Jeremy Needle, Robert Fromont, and Seán Roberts for
his gyfarwyddyd. All faults remain ours.

References

Atkinson, M., S. Kirby, and K. Smith (2015). Speaker input variability does not explain why larger populations have simpler languages. *PloS One* 10(6), e0129463.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). Evaluating online labor markets for experimental research: Amazon.com’s Mechanical Turk. *Political Analysis* 20(3), 351–368.

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

Borji, A. and L. Itti (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on pattern analysis and machine intelligence* 35(1), 185–207.

Brunelli`ere, A. and S. Soto-Faraco (2013). The speakers’ accent shapes the listeners’ phonological predictions during speech perception. *Brain and Language* 125(1), 82–93.

Bucholtz, M. (1999). ‘Why be normal?’: Language and identity practices in a community of nerd girls. *Language in Society* 28(2), 203–223.

Campbell-Kibler, K. (2011). Intersecting variables and perceived sexual orientation in men. *American Speech* 86(1), 52–68.

Chater, N. and C. D. Manning (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Science* 10(7), 335–344.

Cheshire, J. (2002). Sex and gender in variationist research. *Handbook of Language Variation and Change*, 423–443.

Chevrot, J.-P. and P. Foulkes (2013). Introduction: Language acquisition and sociolinguistic variation. *Linguistics* 51(2), 251–254.
Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

Clopper, C. G. and D. B. Pisoni (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics* 32(1), 111–140.

Cohn, A. C., W. H. Ham, and R. J. Podesva (1999). The phonetic realization of singleton-geminate contrasts in three languages of Indonesia. In *Proceedings of ICPhS*, Volume 14, pp. 587–590. Citeseer.

Crump, M. J., J. V. McDonnell, and T. M. Gureckis (2013). Evaluating Amazon’s Mechanical Turk as a tool for experimental behavioral research. *PloS One* 8(3), e57410.

De Vaan, L., R. Schreuder, and R. H. Baayen (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon* 2(1), 1–23.

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

Eckert, P. (2000). *Linguistic variation as social practice: The linguistic construction of identity in Belten High*, Volume 27. Hoboken, NJ: Wiley-Blackwell.

Eckert, P. and S. McConnell-Ginet (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology* 21(1), 461–488.

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology* 1(1), 5–39.

Foulkes, P. and G. Docherty (2006). The social life of phonetics and phonology. *Journal of Phonetics* 34(4), 409–438.

Foulkes, P., G. Docherty, G. Khattab, and M. Yaeger-Dror (2010). Sound judgments: Perception of indexical features in children’s speech. In D. R. Preston and N. Niedzielski (Eds.), *A reader in sociophonetics*, Trends in linguistics 219, pp. 327–356. New York, NY: De Gruyter Mouton.

Foulkes, P. and J. Hay (2015). The emergence of sociophonetic structure. *The Handbook of Language Emergence* 87, 292.
Giles, H., D. M. Taylor, and R. Bourhis (1973). Towards a theory of interpersonal accommodation through language: Some Canadian data. *Language in Society* 2(02), 177–192.

Godden, D. R. and A. D. Baddeley (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology* 66(3), 325–331.

Habick, T. (1991). Burnouts versus rednecks: effects of group membership on the phonemic system. *New ways of analyzing sound change* 185, 212.

Harrington, J., S. Palethorpe, and C. I. Watson (2000). Does the Queen speak the Queen’s English? *Nature* 408(6815), 927.

Hay, J. (2018). Sociophonetics: The role of words, the role of context, and the role of words in context. *Topics in Cognitive Science*.

Hay, J. and K. Drager (2007). Sociophonetics. *Annual Review of Anthropology* 36, 89–103.

Hay, J. and K. Drager (2010). Stuffed toys and speech perception. *Linguistics* 48(4), 865–892.

Hay, J., A. Nolan, and K. Drager (2006). From fush to feesh: Exemplar priming in speech perception. *Linguistic Review* 23(3), 351–379.

Hay, J., R. Podlubny, K. Drager, and M. McAuliffe (2017). Car-talk: Location-specific speech production and perception. *Journal of Phonetics* 65, 94–109.

Hay, J., A. Walker, K. Sanchez, and K. Thompson (2019). Abstract social categories facilitate access to socially skewed words. *PloS one* 14(2), e0210793.

Hudson Kam, C. L. and E. L. Newport (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* 1(2), 151–195.

Jannedy, S., M. Weirich, and J. Brunner (2011). The Effect of Inferences on the Perceptual Categorization of Berlin German Fricatives. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2011)*, 962–965.
Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

Johnson, F. L. and R. Buttny (1982). White listeners’ responses to “sounding black” and “sounding white”: The effects of message content on judgments about language. *Communications Monographs* 49(1), 33–49.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. *Talker variability in speech processing*, 145–165.

Jurafsky, D. (2012). Universals in the semantics of the diminutive. In *Annual Meeting of the Berkeley Linguistics Society*, Volume 19.

Kam, C. L. H. and E. L. Newport (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology* 59(1), 30–66.

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

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

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

Ladegaard, H. J. and D. Bleses (2003). Gender differences in young children’s speech: the acquisition of sociolinguistic competence. *International Journal of Applied Linguistics* 13(2), 222–233.

Langstrof, C. (2014). *Sociophonetic learning in L1 and L2*. Habilitation, University of Freiburg.

Lawson, E., J. Scobbie, and J. Stuart-Smith (2011). The social stratification of tongue shape for post-vocalic /r/ in Scottish English. *Journal of Sociolinguistics* 15(2), 256–268.

Leung, J. H. and J. N. Williams (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning* 62(2), 634–662.

Leung, J. H. and J. N. Williams (2013). Prior linguistic knowledge influences implicit language learning. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, pp. 2866–2871. Austin, TX: Cognitive Science Society.
Lewicki, P., T. Hill, and M. Czyzewska (1992). Nonconscious acquisition of information. *American Psychologist* 47(6), 796.

MacFarlane, A. E. and J. Stuart-Smith (2012). ‘One of them sounds sort of Glasgow Uni-ish’. Social judgements and fine phonetic variation in Glasgow. *Lingua* 122(7), 764–778.

MATLAB (2016). *version 9.0 (R2016a)*. Natick, Massachusetts: The MathWorks Inc.

Maye, J. and D. Weiss (2003). Statistical cues facilitate infants’ discrimination of difficult phonetic contrasts. In *Proceedings of the 27th annual Boston University conference on language development*, Volume 2, pp. 508–518.

Maye, J., J. F. Werker, and L. Gerken (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82(3), B101–B111.

Mendoza-Denton, N. (1996). ‘muy macha’: Gender and ideology in gang-girls’ discourse about makeup. *Ethnos* 61(1-2), 47–63.

Metcalfe, J. (2017). Learning from errors. *Annual review of psychology* 68, 465–489.

Milroy, J. and L. Milroy (1993). Mechanisms of change in urban dialects: The role of class, social network and gender. *International Journal of Applied Linguistics* 3(1), 57–77.

Milroy, L. (1980). *Language and social networks*. Oxford.

Molnar, M., A. Ibáñez-Molina, and M. Carreiras (2015). Interlocutor identity affects language activation in bilinguals. *Journal of Memory and Language* 81, 91–104.

Needle, J. M. and J. B. Pierrehumbert (2018). Gendered associations of English morphology. *Laboratory Phonology* 9(1).

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology* 18(1), 62–85.
Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

Pharao, N., M. Maegaard, J. S. Møller, and T. Kristiansen (2014). Indexical meanings of [s+] among Copenhagen youth: Social perception of a phonetic variant in different prosodic contexts. *Language in Society* 43(01), 1–31.

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

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics* 2(1).

Pierrehumbert, J. B., M. E. Beckman, and D. R. Ladd (2000). Conceptual foundations of phonology as a laboratory science. *Phonological knowledge: Conceptual and empirical issues*, 273–304.

Preston, D. R. (1996). Whaddayaknow?: The modes of folk linguistic awareness. *Language Awareness* 5(1), 40–74.

Qian, T., T. Jaeger, and R. N. Aslin (2014). Implicit learning in a non-stationary environment: Knowing when it’s Groundhog Day. ms, University of Rochester.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Roberts, G. (2017). The linguist’s drosophila: Experiments in language change. *Linguistics Vanguard* 3(1).

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

Sankoff, G. and H. Blondeau (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, 560–588.

Scott-Phillips, T. C. and S. Kirby (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences* 14(9), 411–417.
Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

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

Smith, J., M. Durham, and H. Richards (2013). The social and linguistic in the acquisition of sociolinguistic norms: caregivers, children, and variation. *Linguistics 51*(2), 285–324.

Sneller, B. and G. Roberts (2018). Why some behaviors spread while others don’t: A laboratory simulation of dialect contact. *Cognition 170*, 298–311.

Timmins, C., F. Tweedie, and J. Stuart-Smith (2004). *Accent change in Glaswegian (1997 corpus): Results for consonant variables*. Glasgow, UK: The Glasgow Speech Project, University of Glasgow.

Trechter, S. (1995). Categorical gender myths in Native America: Gender deictics in Lakhota. *Issues in Applied Linguistics 6*(1), 5–22.

Trudgill, P. (1974). *The Social Differentiation of English in Norwich*. Cambridge University Press.

Van Berkum, J. J., D. Van den Brink, C. M. Tesink, M. Kos, and P. Hagoort (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience 20*(4), 580–591.

Walker, A. and J. Hay (2011). Congruence between ‘word age’ and ‘voice age’ facilitates lexical access. *Laboratory Phonology 2*(1), 219–237.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

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

**Appendix: Model fitting**

**Model 1**

The first set of comparisons tested main term interactions. The best model was reported.
The model fitting process is outlined in Table 4. Each model was compared to model id (1.)1. The table spells out the model formulae and reports model fitting statistics, namely the degrees of freedom (df), the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), log likelihood (logLik), and deviance (which is $-2 \times \log \text{likelihood}$). First, we specified a model with no interaction terms (1.1). Then, we tested all relevant interactions (1.2-5). Interactions between robust predictors were considered relevant. Pairwise Chi-square goodness-of-fit tests and the Akaike Information Criterion were used to determine the best fit. When these conflicted, we relied on the p value provided by the Chi-square test.

The second set of comparisons test for the robustness of the main terms. Non-robust main terms were reported in the best model. Model comparisons lend additional support to the Wald confidence intervals showing that these terms (viz. familiarity with conversation partner / item and competitor cue type) did not explain additional variation in the test data. Details can be seen in Table 5: Here, we report pairwise comparisons with model id 2 from Table 4.

The only model that improves on the baseline model (1.1) is (1.2). One interaction is therefore justified in the model. We also refit (1.2) with random slopes for main cue and conversation partner, but these did not improve fit or resulted in a singular fit, indicating overfitting (not included). We report the model estimates with no slopes, using Wald intervals, in Figure 5 as well as in Table 6 below.

The reader will note that two sets of comparisons were made here. First, we consider interactions of main terms. Then we pick the best model and consider the robustness of its main terms.

**Model 2**

Model 2 compares tests with the diminutive (reported above) and plural patterns.

It is a multilevel binomial generalised linear regression model predicting an accurate response in a test trial based on cue type as well as the type of morphological distinction represented in the prompt-response image pair (diminutive or plural). Model comparison revealed that the type of morphological distinction does not explain additional variation in the data (see Table 7), indicating that the plural image pairs are not easier or harder to rely on than the diminutive image pairs (an interaction of condition and pattern was
Rác, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

| id | formula                                                                 | term  | df   | AIC     | BIC     | logLik    | deviance | statistic | Chi.Df | p.value |
|----|--------------------------------------------------------------------------|-------|------|---------|---------|-----------|----------|-----------|--------|---------|
| 1  | 1 + main cue + partner familiar + item familiar + competitor cue + (1 | fit1  | 10.00| 35573.69| 35658.43| -17776.85| 35553.69 |          |        |         |
|    | | familiar + competitor cue + (1 | subject) |       |         |          |          |           |        |         |
| 2  | 1 + main cue * partner familiar + item familiar + competitor cue + trial count (rescaled) + (1 | fit2  | 14.00| 35426.29| 35544.92| -17699.15| 35398.29 | 155.40 | 4.00    | 0.00    |
|    | | subject) |       |         |          |          |           |           |        |         |
| 3  | 1 + main cue * item familiar + partner familiar + competitor cue + (1 | fit3  | 13.00| 35576.69| 35686.84| -17775.34| 35550.69 | 3.00   | 3.00    | 0.39    |
|    | | subject) |       |         |          |          |           |           |        |         |
| 4  | 1 + main cue + item familiar + partner familiar * competitor cue + (1 | fit4  | 13.00| 35575.67| 35685.83| -17774.84| 35549.67 | 4.02   | 3.00    | 0.26    |
|    | | subject) |       |         |          |          |           |           |        |         |
| 5  | 1 + main cue + partner familiar + item familiar * competitor cue + (1 | fit5  | 13.00| 35575.96| 35686.11| -17774.98| 35549.96 | 3.73   | 3.00    | 0.29    |
|    | | subject) |       |         |          |          |           |           |        |         |

Table 4: Model 1: Model comparisons for interactions of main terms
Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

| model id | formula                                                                 | term | df     | AIC      | BIC      | logLik  | deviance | statistic | ChildDF | p.value |
|----------|--------------------------------------------------------------------------|------|--------|----------|----------|---------|----------|-----------|---------|---------|
| 6        | 1 + main cue * partner familiar + competitor cue + trial count (rescaled) + (1 | fit6  | 13.00  | 35425.11 | 35535.26 | -17699.55 | 35399.11 |
| 2        | 1 + main cue * partner familiar + item familiar + competitor cue + trial count (rescaled) + (1 | fit2  | 14.00  | 35426.29 | 35544.92 | -17699.15 | 35398.29 | 0.81 | 1.00 | 0.37 |
| 7        | 1 + main cue * partner familiar + item familiar + trial count (rescaled) + (1 | fit7  | 11.00  | 35424.37 | 35517.58 | -17701.19 | 35402.37 | 4.08 | 3.00 | 0.25 |
| 2        | 1 + main cue * partner familiar + item familiar + competitor cue + trial count (rescaled) + (1 | fit2  | 14.00  | 35426.29 | 35544.92 | -17699.15 | 35398.29 | 0.81 | 1.00 | 0.37 |
| 8        | 1 + main cue * partner familiar + item familiar + competitor cue + (1 | fit8  | 13.00  | 35566.23 | 35676.38 | -17770.12 | 35540.23 | 4.08 | 3.00 | 0.25 |
| 2        | 1 + main cue * partner familiar + item familiar + competitor cue + trial count (rescaled) + (1 | fit2  | 14.00  | 35426.29 | 35544.92 | -17699.15 | 35398.29 | 141.94 | 1.00 | 0.00 |

Table 5: Model 1: Model comparisons for main terms
Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal

| term                                           | estimate | std.error | z value | 2.5% CI  | 97.5% CI |
|------------------------------------------------|----------|-----------|---------|----------|----------|
| (Intercept)                                    | 0.59     | 0.22      | 2.75    | 0.17     | 1.01     |
| main cue = Age                                 | 0.79     | 0.30      | 2.65    | 0.21     | 1.38     |
| main cue = Ethnicity                           | 0.22     | 0.30      | 0.73    | -0.37    | 0.80     |
| main cue = Gender                              | 0.79     | 0.31      | 2.58    | 0.19     | 1.39     |
| familiar conversation partner                  | -0.03    | 0.05      | -0.57   | -0.12    | 0.07     |
| familiar item                                  | -0.02    | 0.03      | -0.91   | -0.07    | 0.03     |
| competitor cue = Age                           | 0.14     | 0.25      | 0.55    | -0.35    | 0.62     |
| competitor cue = View                          | 0.25     | 0.30      | 0.85    | -0.33    | 0.84     |
| competitor cue = Ethnicity                     | -0.32    | 0.33      | -0.96   | -0.97    | 0.33     |
| trial.count (rescaled)                         | -2.42    | 0.19      | -12.86  | -2.79    | -2.05    |
| main cue = Age * familiar conversation partner | 0.16     | 0.07      | 2.40    | 0.03     | 0.29     |
| main cue = Ethnicity * familiar conversation partner | 0.25   | 0.08      | 3.10    | 0.09     | 0.41     |
| main cue = Gender * familiar conversation partner | 0.02 | 0.07      | 0.31    | -0.12    | 0.16     |

Table 6: Model summary for estimates on test results

also tested and consequently discarded, not shown).

**Model 3**

Model 3 tests whether perceptual distance between images explains more variation in the data than conceptual distance (our main and competitor cues). Details can be seen in Table 8.
Table 7: Model 2: Model comparisons for pattern type

| id | formula                      | Df | AIC    | BIC    | logLik | deviance | Chisq | Chi Df | Pr(>Chisq) |
|----|------------------------------|----|--------|--------|--------|----------|-------|--------|------------|
| 9  | 1 + cond + pattern + (1 | subject) | 3.00 | 10750.13 | 10772.13 | -5372.06 | 10744.13 |
| 10 | 1 + cond + (1 | subject) | 4.00 | 10752.12 | 10781.46 | -5372.06 | 10744.12 | 0.01 | 1.00 | 0.94 |

Rácz, Hay & Pierrehumbert (to appear): Not all indexical cues are equal.
Table 8: Model 3: Model comparisons for image distance type

| formula                                                                 | Df | AIC     | BIC     | logLik | deviance | Chisq | Chi Df | Pr(>Chisq) |
|------------------------------------------------------------------------|----|---------|---------|--------|----------|-------|--------|-------------|
| 1 + main distance (rescaled) * competitor distance (rescaled) + (1 | 7.00 | 35443.43 | 35502.74 | -17714.71 | 35429.43 |       |        |             |
|   (1 | training partner image pair)                                         |    |         |         |        |          |       |        |             |
| 1 + main cue * partner familiar + item familiar + competitor cue + trial count (rescaled) + (1 | 14.00 | 35426.29 | 35544.92 | -17699.15 | 35398.29 | 31.13 | 7.00 | 0.00        | (1 | subject)   |