Co-occurrence, extension and social salience: The emergence of indexicality in an artificial language

Aini Li, Gareth Roberts

Department of Linguistics
University of Pennsylvania
Philadelphia, PA, 19104, USA
corresponding author: Gareth Roberts, gareth.roberts@ling.upenn.edu

Abstract

We investigated the emergence of sociolinguistic indexicality using an artificial-language-learning paradigm. Sociolinguistic indexicality involves the association of linguistic variants with nonlinguistic social or contextual features. Any linguistic variant can acquire "constellations" of such indexical meanings, though they also exhibit an ordering, with first-order indices associated with particular speaker groups and higher-order indices targeting stereotypical attributes of those speakers. Much natural-language research has been conducted on this phenomenon, but little experimental work has focused on how indexicality emerges. Here, we present three miniature artificial-language experiments designed to break ground on this question. Results show ready formation of first-order indexicality based on co-occurrence alone, with higher-order indexicality emerging as a result of extension to new speaker groups, modulated by the perceived practical importance of the indexed social feature.

Keywords: Indexicality; enregisterment; sociolinguistics; language variation; social meaning; artificial language learning

1 Introduction

Language conveys not only semantic meaning but also social meaning. For instance, the sentence "I had a can of pop with my tea" not only conveys what the speaker drank with their meal, but also (via the words *pop* and *tea* as opposed to, say, *soda* and *dinner*) carries implications about their background (Dyer, 2007). Individuals may be categorical in their use of a particular variant. For example, someone who drinks "pop" with their "tea" might never use any alternative variants to refer to carbonated drinks or their evening meal. Often, however, linguistic variation involves the use of multiple variants by the same person. A lawyer, for example, is likely to talk rather differently when addressing a court from how they talk in conversation with a client, a colleague, a family member, or a school friend. Such differences in *register* or *style*, whereby linguistic variation is conditioned by context, involves *intra*individual rather than *inter*individual variation (Biber & Conrad, 2019).[1]

However, the two kinds of social variation are very closely related, and a tendency for coherence between intraindividual stylistic variation and interindividual community-level variation, an important component of what is known as "orderly heterogeneity", has long been recognized (Weinreich, Labov, & Herzog, 1968; Guy & Hinskens, 2016; Roberts & Sneller, 2020). The linguistic variants associated with more formal registers, for instance, tend also to be the variants particularly associated with speakers higher on the social scale. Those speakers might use other variants, of course, but typically with lower frequency. Part of the reason for this is that variants can become associated with particular social registers precisely because of who is associated with using them (Agha, 2003). This can have deep historical roots; the association in English of words of French origin with particular registers, for instance, reflects in great part historical perceptions of language and power dating to the Norman Conquest (Hughes, 1990; J. J. Smith, 2013).

This should not be taken to imply that the attachment of social meaning to linguistic forms is generally a static, slow-moving process, however. In fact, it is a rather dynamic one. To take an obvious example: slang words, slurs, swearwords, and euphemisms often undergo rapid replacement within single generations, rather like change in dress (cf. Burridge, 2012; Blank, 1999, and Acerbi, Ghirlanda, & Enquist, 2012, on nonlinguistic fashion). Indeed, long-term patterns like the association of French with higher registers should be seen not as static images but rather as slower changing backgrounds to much more rapidly evolving foregrounds.

A key concept here is *indexicality*, whereby something can be taken to point towards something else, often in a different domain. For instance, words of French origin came to index prestige via their association with the post-1066 ruling class in England. Similarly, linguistic forms associated with Canada (such as the Canadian raising represented in spellings of *about* as *aboot*) are often taken to index not only Canada, but also—via stereotypes of Canadians—niceness and decency. Almost any linguistic variant can acquire indexical social meaning of this kind, though such meaning is not necessarily simple, unitary, or fixed. Rather, linguistic variants are argued to occupy a constantly evolving "field" or "constellation" of related social meanings (Eckert, 2008). Just as a linguistic variant associated with Canadians might acquire further associations with the stereotypically Canadian trait of niceness and decency, this might in turn lead—via stereotypes of what it is to be nice and decent—to new associations with (e.g.) gullibility. This kind of indirectly acquired social meaning is known as *higher-order indexicality* (Silverstein, 2003; Eckert, 2008). First-order indexicality involves the association of a linguistic variant with a group

[1] For the purposes of this paper we will not draw a sharp distinction between style and register. The two are closely connected, with the term register tending to imply a more reified entity. We direct readers to Irvine (2001) for further discussion of this point.

whose members use it. Higher-orders of indexicality arise as new associations are built on this. Register, whereby particular linguistic forms become associated with particular discourse contexts, can be considered a special case of higher-order indexicality (Eckert, 2008).

Theoretical work on indexicality has inspired a large body of empirical work involving naturalistic data, which has shed light on the ways in which people use language to shape and convey their identity and how they acquire existing indexical associations (e.g., Podesva, 2007; Jaffe, 2009; Meyerhoff & Schleef, 2012; Pharao, Maegaard, Møller, & Kristiansen, 2014). An important question concerns *enregisterment*, the process by which (in the case of language) clusters of linguistic and social features come to be reified as cognitive entities beyond their original context of use, including the establishment of indexical meanings (Agha, 2007; Johnstone, 2016). In Agha's (2007) words, this involves "diverse behavioral signs [being ...] functionally reanalyzed as cultural models of action" (p. 55), while Johnstone (2016) referred to a sign as possessing indexicality "by virtue of co-occurring with what it is taken to mean" (p. 633).

But how exactly might co-occurrence lead to the emergence of indexicality, a process also known as enregisterment?[2] Is mere co-occurrence (or, rather, perceived co-occurrence) sufficient? Agha's (2007) use of the term "reanalyzed" is likely important here. Reanalysis requires not only that a particular correspondence is acquired but also that language users are motivated to reinterpret the correspondence as going beyond its current context and to apply it in some new way. Nor is all co-occurrence necessarily equal. Rácz, Hay, and Pierrehumbert (2017), for instance, exposed participants to an artificial language with morphological variation. When this variation was conditioned by gender, it was readily learnable. When it was conditioned instead by spatial orientation, it was not. Along related lines, Sneller and Roberts (2018) conducted an experiment in which participants learned an artificial language with dialectal variation and used it to interact with each other in a computer game. Variants spread between dialects and did so significantly more readily when they indexed traits that played a practical role in the game.

Both these studies (along with non-experimental sociolinguistic work; e.g., Foulkes, 2010; Levon & Fox, 2014; Llamas, Watt, & MacFarlane, 2016) point towards a role for *salience* in the process of establishing indexical relationships. Indeed, Rácz et al. (2017) explicitly discussed the difference between gender and spatial orientation in terms of "social salience" or the robust "social interpretability" (p. 4) of one trait compared with the other. (See also Needle & Pierrehumbert, 2018, for a natural-language experimental study making reference to gender as a "salient" social category.) Instead of salience per se, Sneller and Roberts (2018) discussed the "practical social relevance" or "social value" of nonlinguistic traits, which they manipulated by varying the role of the trait in the game that their participants played. This recalls broader and long-standing literature on the role of *personal involvement* in perceptual salience more broadly (Borgida & Howard-Pitney, 1983; Fischer, Meyers, Cummins, Gibson, & Baker, 2020). A way of summarizing this is to say that language users are likely to be influenced in forming indexical associations by the relative salience of the traits involved, with the social value of a trait in a given context playing a role in determining this. There are different kinds of social value, however. A trait may have social value relative to other traits. For example, physical toughness may be more important than intellectual prowess in a street gang; at a university the reverse is likely to be true (Denton, Kruschke, & Erickson, 2008; Kraljic, Brennan, & Samuel, 2008). But there is also the social value of the *index* itself, distinct from the trait it is indexing. If toughness is already well indexed, then a

[2] Strictly speaking, enregisterment and the emergence of indexicality are not treated by all researchers as quite identical; in particular, enregisterment implies the creation of a somewhat reified indexical entity; for the sake of brevity, however, we will in this paper treat enregisterment and the process of establishing indexicality as the same thing for our purposes.

new index is unlikely to be needed unless it allows new social distinctions to be made.

The research reviewed here allows us to make some predictions. First, based on Agha (2007), we can predict that linguistic enregisterment requires covariance between linguistic variation and social traits. Second, based on Sneller and Roberts (2018), we can predict that it will occur when indexicality serves a social purpose (such as allowing a new social distinction to be made). Third, based on Rácz et al. (2017), we can predict that indexicality will tend to attach to the most socially salient available traits. In this paper, we present a series of experiments designed to test these predictions. In doing so we employ an artificial-language paradigm. An important advantage of such approaches is that they allow considerably greater control than is possible in non-experimental research on natural language (Roberts & Sneller, 2020; Roberts, 2017), and several such studies have taken this approach in explicitly testing sociolinguistic hypotheses (e.g., Sneller & Roberts, 2018; Wade & Roberts, 2020; Lai, Rácz, & Roberts, 2020; Rácz et al., 2017, 2017). Related to this work, and to the study we present here, is a strand of work concerning the acquisition of grammatically unpredictable variation in artificial languages. This research suggests that learners will tend to condition such variation on available linguistic or social cues (K. Smith & Wonnacott, 2010; Hudson Kam & Newport, 2005; Vihman, Nelson, & Kirby, 2018; Samara, Smith, Brown, & Wonnacott, 2017).

Little such work has shed direct light on enregisterment, however. The study we present here is intended to lay groundwork for this question. The study involved exposing participants to an artificial language in which grammatically unpredictable variation co-occurred reliably with two different non-linguistic traits: alien species and clothing. After exposing participants to this pattern of co-occurrence, we measured what associations participants had acquired between aliens, outfits, and suffixes.

We conducted three experiments in total. In **Experiment 1**, we investigated whether participants would primarily acquire indexical associations between suffixes and speakers or between suffixes and the speakers' clothing (predicting that, for the reasons given above, it would be the former; cf. Rácz et al., 2017).[3] In **Experiment 2**, we introduced a new species of alien in the association test phase, who wore the same clothing as the other aliens; this allowed us to test whether participants had, alongside primary associations with speaker group, formed secondary associations with clothing that they would extend to new contexts. Finally, in **Experiment 3**, we manipulated the social importance of the clothing to investigate if this strengthened the indexical associations participants formed with clothing.

In conditions in which we did not give clothing extra social importance (as in all conditions of Experiments 1 and 2), we had an expectation that the two variables would behave differently. In particular, we predicted that clothing would be more likely than species to be backgrounded, in line with known tendencies for humans to perceive faces as more visually interesting and worthy of attention than bodies (Kwon, Setoodehnia, Baek, Luck, & Oakes, 2016; Wang, Chandler, & Le Callet, 2010; Isola, Xiao, Parikh, Torralba, & Oliva, 2013), and that participants would be more likely to treat species as indexically relevant. This is also in line with patterns of indexicality, whereby first-order indexicality tends to involve speaker group and second-order indexicality tends to be based on social attributes of speaker group (Eckert, 2008; Silverstein, 2003). We anticipated that participants would be more likely to perceive species as an indicator of speaker group than clothing, which is a more "alienable" trait that can be exchanged more easily by speakers (see Sneller & Roberts, 2018, for further discussion of alienability in the context of indexicality). For these reasons we expected species to act, by default, as a more visually and socially salient trait than clothing.

[3] Before conducting Experiment 1, we also conducted a shorter version of the same experiment, whose results were very similar to those of Experiment 1 and which we include in Appendix B.

These two non-linguistic variables were chosen, first, because they are both rather easily operationalized and depicted and, moreover, can be depicted jointly in the same area of the image. That is, if a participant is looking at the alien, they are also looking at the alien's clothing, and vice versa. Second, clothing has the advantage that it can potentially act as a shorthand for a range of other non-linguistic factors such as profession, gang, or context (as with the lawyer example given above) that could have implications for style or register and which would be harder to depict in the same way. It also allowed us to relatively easily manipulate its social significance, as in Experiment 3.

2 Experiment 1: Learning the variation

2.1 Experiment overview

In Experiment 1 we exposed participants to a reliable three-way relationship between linguistic form (specifically plural endings), speaker group (alien species), and cultural trait (clothing) and measured what associations they formed from this exposure (cf. Lai et al., 2020, who employed a very similar task structure). The goal was to test, first, whether participants would form robust associations between linguistic and non-linguistic traits simply as a result of their exposure and, second, whether they would preferentially form associations between the linguistic forms and the speaker groups, between the linguistic forms and the cultural traits, or to a roughly equal extent with both. Our strong expectation, following other work (Rácz et al., 2017; Sneller & Roberts, 2018), was that participants would indeed form associations between linguistic forms and non-linguistic traits and that they would form stronger associations with speaker group (though the magnitude of the preference was not clear). We tested this by manipulating whether, in the test phase, the aliens continued to be seen in the same outfits they had been wearing in the exposure phase (henceforth the *Nonflipped condition*). In the condition in which the aliens switched outfits (henceforth the *Flipped condition*), would participants make linguistic associations based on species or outfit?

2.2 Method

2.3 Participants

A total of 60 participants, recruited through Prolific, completed Experiment 1 in return for \$6. After excluding participants ($N = 4$) whose duration was below the 2.5% quantile or above the 97.5% quantile, we analyzed data from the remaining 56 participants. Of these participants, aged 18–52 ($median = 23$), 23 were female and 33 were male. There were 30 participants in the Flipped condition and 26 participants in the Nonflipped condition.

2.3.1 Alien language

The "alien" language that participants were trained on contained ten noun stems, as shown in Table 1. The ten word stems were randomly generated by combining CV or CVC syllables using the consonants /k, g, q, h, w, j, l, t, s, m, r, p/ and the vowels /a, e, i, o, u/. Each of the ten nouns was used to refer to one of ten objects. For instance, the word "kabuq" might be used to refer to the green vegetable illustrated in Figure 1. There were also two plural endings, *-dem* and *-gok*, each of which could be affixed to any noun stem, their distribution being entirely determined by the alien speaker using the word and not by any feature of the noun or the object it referred to. Images of all the object stimuli used in the current study can be found in Figure 17, Appendix A.

Table 1: Stem words in the Alien language

kabuq, bupod, hasot, wejun, kenig, tulimur, petilet, ropuko, luragur, gunawul

Figure 1: Example referent with alien-language label

2.3.2 Aliens and outfits

Two alien species were presented as users of the language: the *Nulus* and the *Gilis*. As illustrated in Figure 2, these two alien species differed greatly in their appearance. To avoid the group-level linguistic variation being attributed to individual-level associations, four individuals were further designed respectively for each alien species. In addition, the aliens were presented as wearing two different ceremonial outfits: a black suit-like outfit and a blue cloak-like outfit (Figure 2).

(a) (b) (c) (d)

Figure 2: Nulus (a, b) and Gilis (c, d) in different outfits.

2.4 Procedure

As illustrated in Figure 3, there were six phases in four types in Experiment 1 (presented in the following order): a familiarization phase, two training phases, two memory test phases, and an association test phase. These are described below.

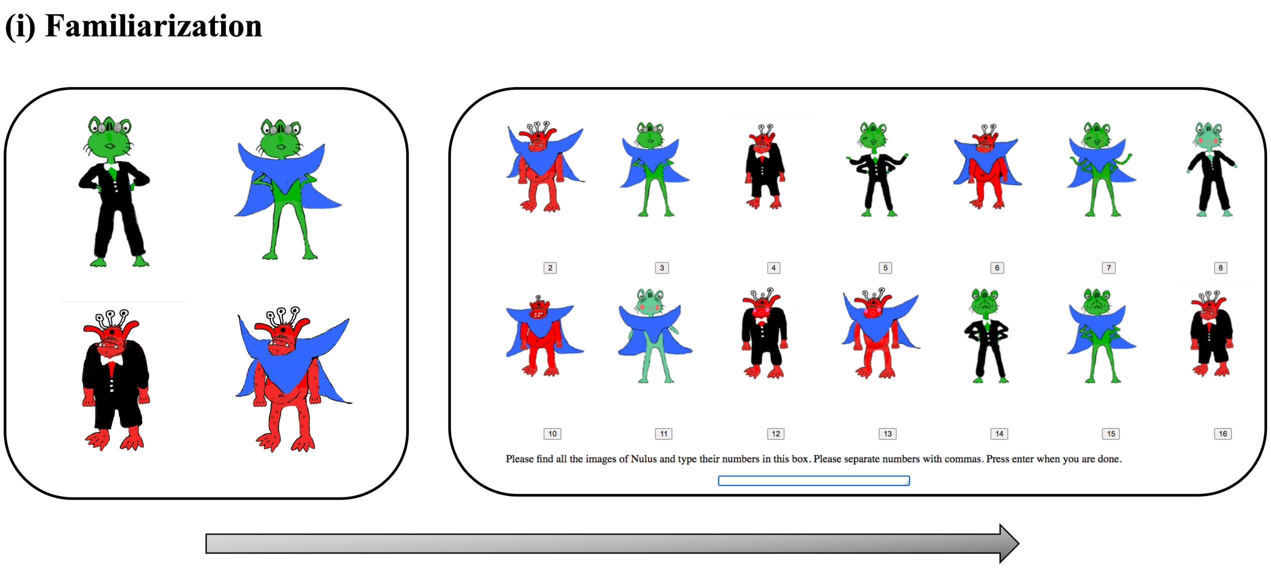
The familiarization phase: The experiment started by introducing the two alien species and their ceremonial outfits. To ensure that participants were fully aware of both, alien species and outfits were presented simultaneously. First, participants saw a screen with four images of aliens (Figure 3(i), left panel): a Nulu wearing the black outfit, the same Nulu wearing the blue outfit, a Gili wearing the black outfit, and the same Gili wearing the blue outfit. Each image was labeled with the words "Nulu/Gili wearing outfit one/two". To ensure participants understood the difference between the species and outfits, this was followed by a grouping exercise. To facilitate the grouping activity, participants were first shown images of one alien wearing both outfits (top-middle panel of Figure 3(i)). Then they saw a screen of 16 aliens labeled with numbers from one to 16 (four different Nulus + four different Gilis $\times$ two outfits interspersed with each other) and were asked to select all the aliens from the species they had just seen by typing their numbers into a text box. After they entered the numbers, a NEXT button would appear and participants could proceed to the next trial by clicking on it. The same grouping task was then repeated for the other species (i.e., the same Gili wearing both outfits were shown and participants needed to identify all the Gilis afterwards) and for each of the two outfits separately (i.e., one Nulu and one Gili both wearing the black or the blue outfit and then participants had to pick out all the right black/blue-outfit-wearing aliens).

Training phases: The training phase began immediately after the familiarization phase. During this phase, participants were instructed to "try to learn what the alien words are for the different objects and how the language works". This phase involved two kinds of trials (Figure 3(ii), bottom-left panel): *passive exposure trials* and *forced-choice trials*. In passive exposure trials, a singular or plural word was presented paired with an image of the object(s) it referred to. In each case, the word was depicted as being spoken by an alien wearing one of the two outfits. Each outfit and plural suffix was paired 100% reliably with one of the two alien species; that is, while participants had seen both species wearing both outfits in the Familiarization phase, one outfit was now seen only on Gilis while the other was seen only on Nulus. The assignment of outfits and suffixes to alien species was counterbalanced between participants. Passive exposure trials were not timed and participants could proceed to the next trial by clicking on a NEXT button.

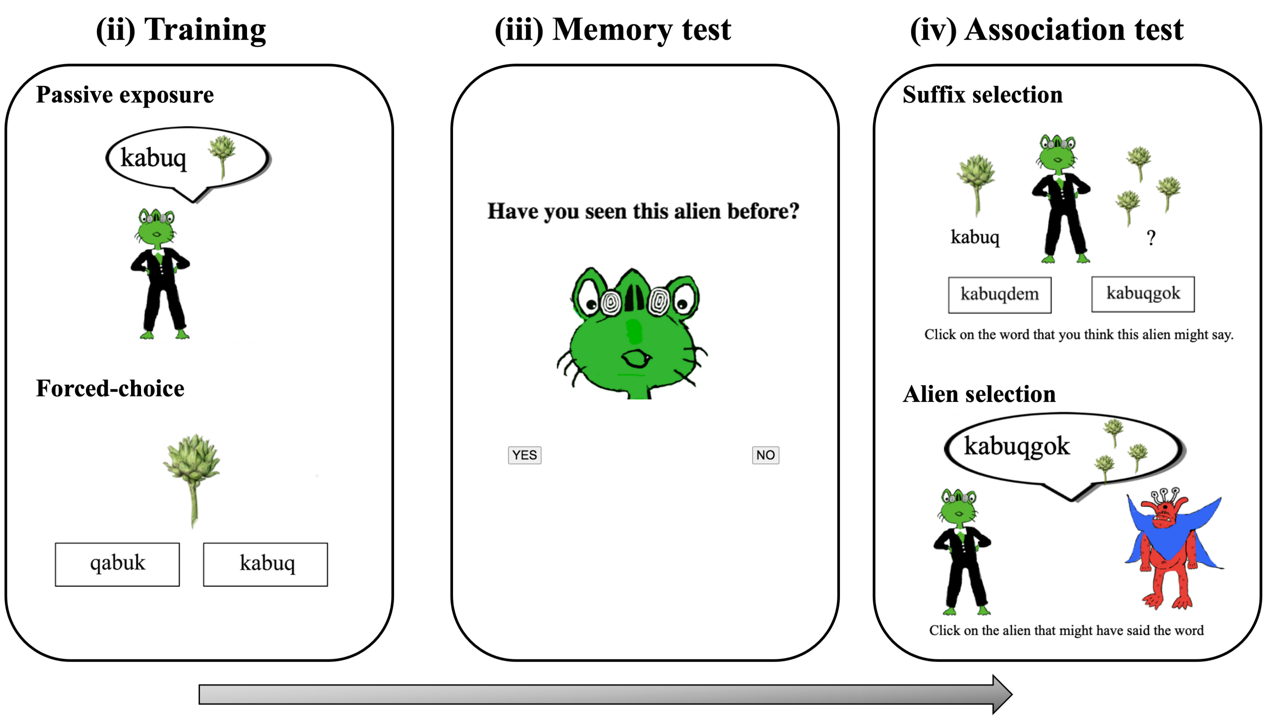
In forced-choice trials participants were shown an image and asked to select an alien word to go with it. In every case the correct word was presented along with a foil word generated by swapping two segments of the correct word for singulars (e.g., kabuq vs. kaqub). For plurals, the swapped segments were always in the suffix (e.g., kabuqgok vs. kabuqkog). This was to ensure that participants would attend to the suffixes as well as the stems. In forced-choice trials, participants had to choose correctly in order to proceed. If the wrong word was chosen, they were told so and asked to try again. Participants were trained on 20 alien words (10 singular and 10 plural) in total. Every set of four passive exposure trials was followed by four forced choice trials on the same four words that the participants had just been exposed to. The order of trials in each such block was randomized. In total, participants went through 10 words $\times$ 2 forms (singular and plural) $\times$ 2 alien species $\times$ 2 trial types (exposure and forced choice) $\times$ 2 repetitions = 160 trials.

There were two training phases, with the second training phase being identical to the first training phase except that it had no forced-choice trials.

The memory test: After each training phase, participants were presented with a total of 12 memory test trials (Figure 3(iii), bottom-middle panel), with three trials each on isolated words, objects, aliens (without outfits) and outfits (not on aliens). Some of the words, aliens, and outfits had been encountered during the training phase; others had not. Participants were instructed to answer yes or no on each screen by pressing on the corresponding buttons to indicate whether or not they recalled seeing the word or im-



(a)



(b)

Figure 3: Structure of Experiment 1

age before. Images of aliens depicted only alien heads so as to separate aliens and outfits in a relatively natural way. Recognition memory tasks of this kind are quite commonly used in artificial-language learning experiments (among other paradigms) for the purposes of checking participants' attention and what they have learned (e.g., Dahan & Brent, 1999; Merkx, Rastle, & Davis, 2011; Havas, Waris, Vaquero, Rodríguez-Fornells, & Laine, 2015). Their purpose in our experiment was (in conjunction with the familiarization phase) also to help reduce the possibility that participants might ignore some of the stimuli (e.g., the outfits). By having participants engage in tasks that required attention to what stimuli they had seen, we hoped to reinforce the importance of the different kinds of stimuli and encourage attention to how they varied. With this in mind, participants were told in the instructions to expect such tests.

There were two memory tests in total, and the second memory test (Memory test 2) was of the same length as the first memory (Memory test 1) but included a different subset of words and objects.

The association test: After completing the memory test, participants began the association test, which was also the final phase of the experiment (Figure 3(iv), bottom-right panel). The aim of this phase was to evaluate the extent to which participants had established associations between alien groups, outfits and plural suffixes. Following Lai et al. (2020), it involved two tasks: a *suffix-selection task* and an *alien-selection task*. No feedback was provided on either task, and trials during this phase proceeded automatically once a choice had been made. Participants were not able to go back to previous trials and change their answers.

In *suffix-selection* trials, participants had to select from a choice of two plural word forms to go with a single alien. These trials were like forced-choice trials in the training phase except in two respects: First, there was an alien present; second, the stems of the optional answers were always identical while the suffixes were always different. In each trial, participants were presented with an alien wearing one of the two outfits. On its left was an image of an object and the alien word that referred to it. On its right was an image of the same object repeated three times accompanied by a question mark. Beneath the alien were two candidate plural forms (in counterbalanced order). The participant's task was to choose the plural form they thought the alien was most likely to use. There were again 40 trials in total (10 suffixed words $\times$ 2 species $\times$ 2 repetitions = 40 trials).

In *alien-selection* trials, by contrast, participants were presented with one plural word form (paired with the object it referred to) and asked to select from two aliens (in counterbalanced order), one from each species and each wearing a different outfit. There were 40 trials in total (10 suffixed words $\times$ 2 species $\times$ 2 repetitions = 40 trials).

2.4.1 Experimental conditions

Participants were randomly assigned to two between-subjects conditions: the *Flipped* condition and the *Nonflipped* condition. The two conditions differed with respect to which outfits were worn by which species in the association test phase. In the training phase, as stated above, each outfit was paired 100% reliably with one of the two alien species. In the *Nonflipped condition*, this pairing was continued in the association test phase. In the *Flipped condition* the pairing was 100% reversed, such that the alien species that had worn one outfit in the training phase now wore the other outfit, and vice versa (Figure 4). Aliens and outfits were counterbalanced across different conditions.

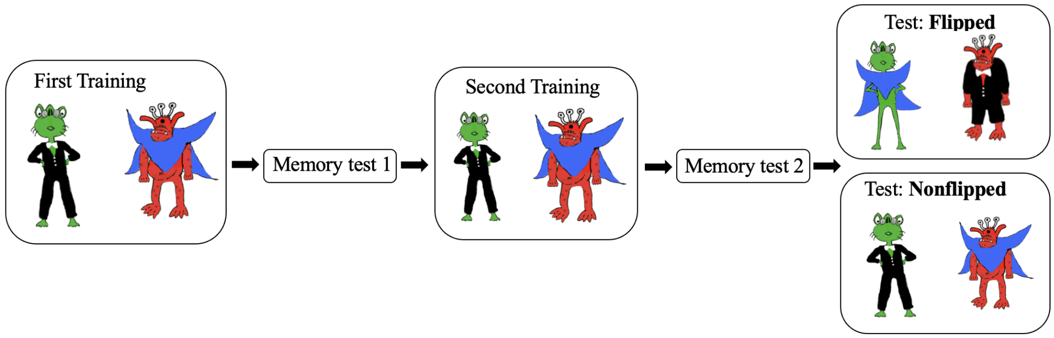


Figure 4: Conditions of Experiment 1

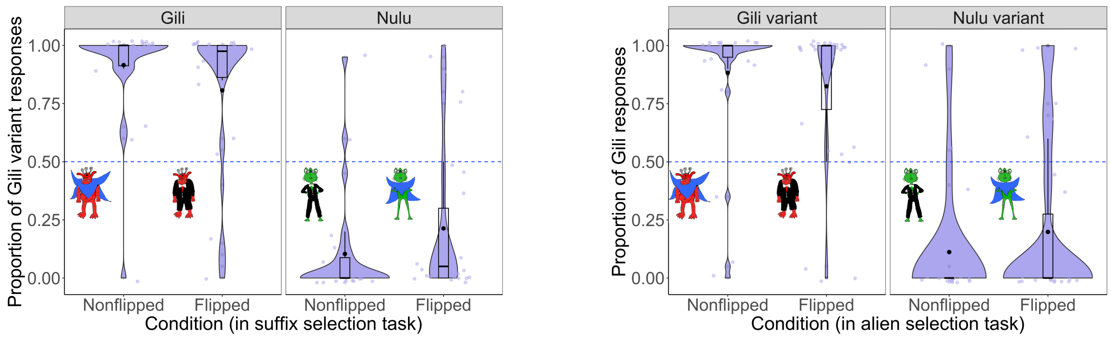
2.4.2 Predictions

Assuming that participants formed associations with the two plural forms based on the training data, there were three main possibilities for the results of the Association phase. First, participants might overwhelmingly associate plural forms with alien species, not outfits. The second possibility was that participants would overwhelmingly associate plural forms with outfits as opposed to aliens. The third possibility was that participants would associate the plural forms with both the aliens and the outfits. All three possibilities predict the same for the Nonflipped condition: Participants' responses should line up with the training data. The prediction for the Flipped condition differs for all three possibilities, however. If participants mainly associated plural endings with alien species, then participants should pick the same aliens for the same suffixes (and vice versa) across both conditions. If, by contrast, they mainly associated endings with outfits, then they should reverse their alien species choices in the Flipped condition relative to the Nonflipped condition. If, on the other hand, they formed associations with both aliens and outfits, then selections in the Flipped condition should be more variable than in the Nonflipped condition and (if associations with aliens and outfits were similarly strong) should be close to chance level.

Based on earlier literature (e.g., Rácz et al., 2017), we predicted that the second possibility was unlikely. Given that species is likely to be more socially salient than clothing (see discussion above), it would be surprising if participants overwhelmingly associated the forms with the clothing rather than the aliens. This is also consistent with how indices are observed to arise in the real world: Second-order indices tend to emerge as secondary associations for linguistic variants already associated with a speaker group (Silverstein, 2003; Eckert, 2008; Sneller & Roberts, 2018). However, even if (as expected) participants formed a strong primary association between linguistic forms and alien species, they might also develop a significant *secondary* association with the outfits—possibility three—which should lead to greater variability in responses in the Flipped condition.

2.5 Results for Experiment 1

Analyses were conducted using the R Statistical environment (R Core Team, 2017). Generalized mixed-effects logistic regression was conducted using the glmer function from the `lme4` package, version 1.1-27.1 (Bates, Mächler, Bolker, & Walker, 2014). Post-hoc pairwise comparisons were extracted using the estimated marginal means (emmeans) package version 1.5.5-1 (Lenth & Lenth, 2018), which adjusts the p-values using the Tukey method to correct for multiple comparisons. Plots were created using ggplot2 package version 3.3.5 (Wickham, 2016). Model coefficients were plotted using broom package version



(a) (b)

Figure 5: Results for Experiment 1 (black dots indicate means; purple dots indicate participant means; blue dotted line indicates chance level): (a): suffix selection task; (b): alien selection task

1.0.2. Data and analysis scripts are available at https://bit.ly/3jKvZDT.

Mean completion time was 24.22 min ($sd = 9.04$). Mean recall rate in the memory tests was 67% ($sd = 6\%$). Table 2 summarizes participants' responses in both suffix- and alien selection tasks. As can be seen, during the test phase, even though participants strongly associated the Gili variant with the Gili aliens, they tended to make more switches/errors in the Flipped condition.

Table 2: Confusion matrix for all responses in both tasks

		Response	
		Nonflipped (520 responses in total)	Flipped (600 responses in total)
		Gili variant	Gili variant
Stimulus	Gili alien	476 (91.5%)	484 (80.7%)
	Nulu alien	54 (10.4%)	128 (21.3%)
		Gili alien	Gili alien
	Gili variant	459 (88.3%)	495(82.5%)
	Nulu variant	58 (11.2%)	119 (19.8%)

Figure 5 further shows the aggregate results for suffix selection and alien selection in Experiment 1. In both the suffix- and the alien-selection tasks, participants strongly associated suffixes with alien species, consistent with the first possibility described above (i.e., that participants would overwhelmingly associate plural forms with alien species rather than outfits). There was, however, more variability in the suffix selection task than in the alien selection task (sds = 0.4, 0.36 respectively).

Mixed-effects logistic regression models were fit separately for the two tasks, with Response (suffix selection: Gili variant;[4] Alien selection: Gili species) as the dependent variable, Condition (Nonflipped as the reference level), Stimuli (Gili and the Gili Outfit as the reference level in suffix selection, and the

[4]In what follows, the term "Gili variant" or "Gili suffix" refers to the suffix that participants observed Gilis using in the training phase, and the "Gili outfit" refers to the outfit that participants observed Gilis wearing in training, regardless of what they observed or chose during the association test.

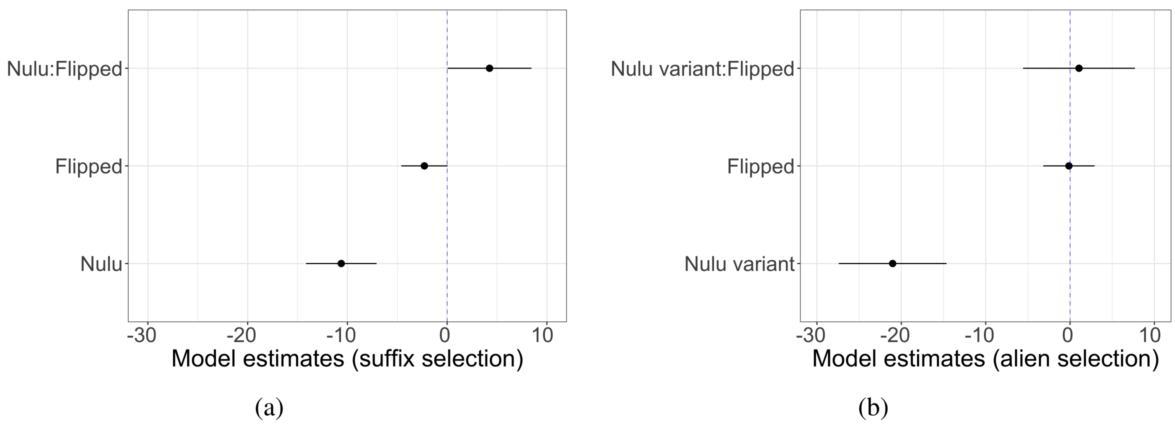


Figure 6: Plot of regression coefficients in suffix selection (a) and alien selection (b) in Experiment 1. Error bars show 95% confidence intervals.

Gili suffix as the reference level in alien selection) and their interaction as independent variables. In addition, we included a by-Word random intercept, a by-Participant random intercept and a by-Participant random slope for Stimuli.[5] Because aliens were always depicted wearing outfits, alien selection can always be modeled either in terms of which alien was selected or which outfit was selected. In what follows we use the terms *outfit stimulus*, *alien stimulus outfit response*, and *alien response* as appropriate.

Model results revealed that for SUFFIX SELECTION there was a significant effect of Stimulus ($\beta = -10.62$, $p < 0.001$, for alien stimulus and $\beta = -10.52$, $p < 0.001$, for outfit stimulus) and a significant effect of Condition ($\beta = -2.28$, $p = 0.05$, for alien stimulus and $\beta = -8.42$, $p < 0.001$, for outfit stimulus). In other words, participants had a strong tendency to match suffixes with the alien species they had seen using them in training, though there was a little more variability in the Flipped condition, suggesting some secondary association with outfits, as indicated by a significant interaction between Stimulus and Condition ($\beta = 4.25$, $p = 0.05$, for alien stimulus and $\beta = 16.82$, $p < 0.001$, for outfit stimulus) in SUFFIX SELECTION. The pattern for ALIEN SELECTION was slightly different. There was a significant effect of Stimulus ($\beta = -21.03$, $p < 0.001$, for alien and $\beta = -21.00$, $p < 0.001$ for outfit responses). But Condition was not significant for alien responses ($\beta = -0.14$, $p = 0.93$). There was, however, a main effect of Condition for outfit responses ($\beta = -19$, $p < 0.001$). Similarly, no significant interaction was found between Stimulus and Condition for alien responses ($\beta = 1.06$, $p = 0.75$) but there was a significant interaction for outfit responses ($\beta = 40.95$, $p < 0.001$). The model formula and the complete model output for Experiment 1 can be found in Appendix D (Table 5). Details of post-hoc comparisons in Experiment 1 can be further found in Table 6 and Table 7. Model predictions are further plotted below in Figure 6.

Taken together, these results suggest that participants overwhelmingly associated suffixes with the aliens who used them, regardless of outfit. But this effect was starker for the alien selection task than for

[5] For the random effects structure, we began with a maximal structure that included by-Word and by-Participant random intercepts, and by-word random slopes for condition, stimuli, and their interaction. However, the model did not converge and returned singular fit errors. We therefore reduced the model to include the largest set of random effects that would converge. The selection process was conducted based on likelihood ratio tests (LRTs) and proceeded backward in a stepwise manner by removing each random effect one at a time until the model converged. First, interaction terms were removed in descending order of complexity; then, random effects that did not significantly improve model fit were removed, starting with those that captured the least variance.

the suffix selection task, as suggested by the post-hoc pairwise comparisons.

2.6 Discussion of Experiment 1

Participants across conditions strongly associated suffixes with the aliens they had seen use them in the training phase, regardless of outfit. However, the interaction effect in the suffix selection task suggested that there might have been some tendency to also make a secondary association with outfits.

The tendency to strongly associate linguistic variants with speaker groups (i.e., alien species) rather than cultural traits (i.e., alien outfits) related to speaker groups is likely driven by relative social—and to some extent visual—salience, as discussed above. That is, by default we would expect participants to find the alien faces more visually interesting (Kwon et al., 2016; Wang et al., 2010) and more socially relevant for the purpose of establishing first-order indexicality (Rácz et al., 2017; Sneller & Roberts, 2018; Eckert, 2008). The results of Experiment 1 established a basis for our second experiment, in which we investigated extension to new speakers via the less strongly associated trait (clothing) and our third experiment, in which we attempted to make progress in investigating the nature of social salience in this context by manipulating the social *importance* of the clothing.

The difference between the two tasks in the Association phase of Experiment 1 (i.e., suffix selection and alien selection) is intriguing, but has parallels in earlier work by Lai et al. (2020), who saw a similar effect, though with more variation in training distributions than in our experiment. The point is that asking, "What would this individual most likely say?" is not the same as asking, "Which individual would most likely say x?" This is in part because the relationship between language and the world is not one-to-one. An analogy can be made with natural language variation here: If asked, "Where do people use the word 'brolly'?", a participant with good knowledge of regional vocabulary is highly likely to suggest the UK. But, in answer to the question, "What do British people call a handheld device for keeping the rain off?" answers would likely be more variable; after all, British people do often use the word "umbrella". It is possible that something similar occurred in the experiment: Even though the associations in training were 100% reliable, participants did see both forms, and to the extent that they imagined within-species variation, might have been biased in how exactly they ordered associations in their minds.

Along similar, though not identical, lines Ma and Komarova (2019) directly compared Object–Label and Label–Object learning (which differ in terms of what is presented first) and found that they had different consequences: Learners who were exposed to the object rather than the label (i.e., Object–Label learning) were better at learning from an inconsistent source. They argued that Object–Label learning and Label–Object learning may be computationally different, with the former involving more frequency boosting (resulting in overmatching) and the latter involving more undermatching. With respect to our study, we should not assume that predicting aliens from suffixes (i.e. in alien selection) and predicting suffixes from aliens (i.e., in suffix selection) are identical tasks; in fact, they may involve different computations that shift the focus of the test, and which may affect the extent to which features of a referent, rather than the referent as a whole, are processed during learning. In this regard, the relative visual salience of the aliens in our experiment compared with the suffixes might also have played a role. Learning is never about a simple "association" between objects and labels. As suggested by Nixon (2020), associations between cues and outcomes are asymmetrical. When multiple cues are present, stronger associations are likely to be established for more salient cues than for less salient cues.

Yet another alternative possibility is simply that, as a memorization task, alien selection was easier for participants than suffix selection because the alien selection trials more closely resembled the training trials. This could have led to more confusion and thus more variability in suffix selection.

Such questions aside, the overall pattern of results suggests that participants strongly associated the plural endings primarily with the alien species rather than with their outfits. Experiment 1 thus establishes a base expectation for further experimentation. Participants had a strong tendency to associate plural endings with speaker group (i.e., alien species), though there was some indication (primarily in suffix selection) that they might have formed a secondary association with clothing. Given this, a key remaining question is how participants might extend what they have learned to new language users and, in particular, whether they would do so on the basis of clothing. Investigating this was the purpose of Experiment 2.

3 Experiment 2: Encountering new aliens

Experiment 2 resembled Experiment 1 except that we introduced a new alien species in the association test phase, and outfits were not flipped on established aliens. We then examined whether participants would associate plural endings with the new species based on outfits.

3.1 Method

3.1.1 Participants

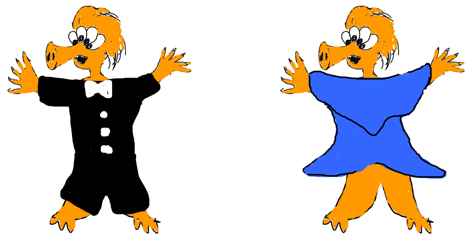
59 participants were recruited through Prolific in return for \$6 each. After excluding participants ($N = 4$) whose duration was below the 2.5% quantile or above the 97.5% quantile of all participants, data from the remaining 55 participants were further analyzed. Of these participants, aged 18–54 ($median = 23$), 26 identified as female, 28 as male, and one as "other".

3.1.2 Procedure

Experiment 2 worked like Experiment 1 except that a new alien species, who had not been previously seen by participants in the training phase, was gradually introduced during the association test phase (Figure 7). There were 64 trials in the association test phase. Participants were informed at the beginning of the association test phase that they might see some new aliens that they have not seen before. The new aliens were introduced gradually. Participants did not encounter new aliens in the first 16 trials but saw only Nulus and Gilis as in Experiment 1. However, the following 14 trials included two trials with new aliens (randomly ordered), and the proportion with new aliens increased from then on, with four in the following 12 trials and 14 in the final 22 trials. Suffix-selection and alien-selection trials occurred equally often overall for each alien. Trials with Nulus or Gilis were the same as in earlier experiments. Trials with new aliens were similar except that, for the alien-selection task, the two aliens were of the same species and differed with respect to outfit only.

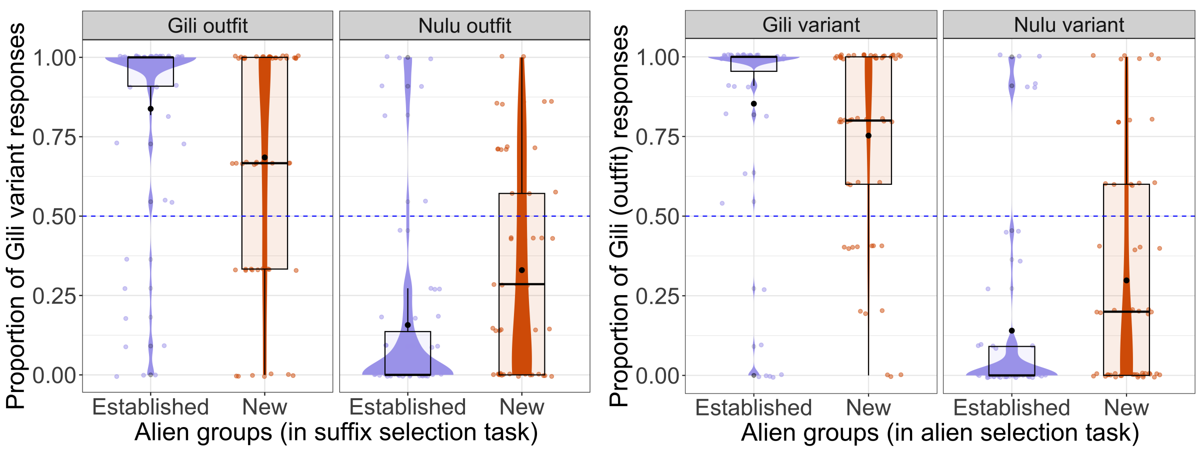
3.2 Results for Experiment 2

Mean completion time was 25.29 min ($sd = 13$). The mean recall rate for the memory test was 67% ($sd = 6\%$). Figure 8 shows results for both new aliens and established aliens in the suffix and alien selection tasks. As can be seen, for established aliens, participants associated the Gili variant strongly with Gilis. For new aliens, however, participants associated the Gili variant with whatever alien was wearing the Gili outfit. As can be seen in Figure 8a, there was nonetheless considerably greater variability for the



(a) (b)

Figure 7: (a) the new alien wearing the black outfit; (b) the new alien wearing the blue outfit



(a) (b)

Figure 8: Results for Experiment 2 (black dots indicate means; purple/orange dots indicate participant means; blue dotted line indicates chance level): (a): suffix selection task; (b): alien selection task

new aliens than for the established aliens. For instance, the standard deviation for Gili variant responses to Gili outfit for New aliens was 0.47, compared with 0.37 for Established aliens: $F = 1.60, p < 0.001$. This suggests that participants had weaker associations for new aliens than for established aliens and that, while they associated linguistic variants with clothing, this was a less certain association than with species. We found similar patterns for the alien selection task (Figure 8b). Variability in responses to New aliens ($sd = 0.43$) was greater than for Established ones ($sd = 0.36$).

Mixed-effects logistic regression models were configured separately[6] for established aliens and new aliens in the suffix selection task, with Response as the dependent variable, Stimulus (Gili Outfit as the reference level) the independent variable. The random effects included by Participant and by Word random intercepts, Stimuli by Participant random slope. Due to counterbalancing, participants were

[6]The number of data points for new aliens was just half of the number of data points for established aliens. Given that the data was thus heavily skewed, we decided to model results for established and new aliens separately rather than collapse the two and include alien group as a predictor in the model, which was not necessary for our main question.

assigned to different versions of the experiment. Therefore, to further capture this variance, Version (i.e., which version participants were exposed to) by Participant was also included as a random slope. According to the model results, for new aliens, there was a significant effect of Stimulus ($\beta = -2.96$, $p < 0.001$), suggesting that participants were significantly less likely to choose the Gili variant for the new alien species when it was wearing a Nulu outfit. Similarly, for established aliens, there was an effect of Stimulus ($\beta = -13.54$, $p < 0.001$). This suggests that for the established aliens, participants were significantly less likely to associate suffix variants with the alien groups that they did not see associated with during the training phase.

We used equivalent models for the alien selection task and found a main effect of Stimulus for both new and established aliens ($\beta = -1.98$, $p < 0.001$; $\beta = -3.58$, $p < 0.001$, respectively). Details of the model formula and the complete model output for Experiment 2 can be found in Appendix D (Table 8).

3.3 Discussion of Experiment 2

Results from Experiment 2 suggest that, while participants acquired strong first-order associations between suffixes and alien species in all experiments, they extended these associations via clothing to previously unencountered aliens in Experiment 2, implying that they had established a latent secondary association with clothing that revealed itself when presented with new aliens.

Participants' responses to new aliens, however, showed much greater variability than for established aliens, suggesting that the associations they had formed between linguistic variables and clothing were not as strong as those they had formed between linguistic variables and species. So far we have discussed this result in terms of the relative *salience* of species and clothing, arguing that this may reflect biases both in visual attention preferences (Kwon et al., 2016; Wang et al., 2010; Isola et al., 2013) and in the social relevance to participants of speaker group vs. clothing. However, we did not manipulate any of these factors, and establishing more clearly the cognitive basis for the apparent salience effects we have observed should be an important focus for future work.

As a first step, we might consider the question of *social importance* in particular. If this is at the root of participants' behavior, then we might expect that associations with clothing should be strengthened if clothing is given more social importance. Earlier artificial-language studies have in fact suggested that attaching practical social value to a cultural trait can have dramatic effects on how participants treat it (e.g., Sneller & Roberts, 2018). In the case of this experiment, the most obvious place to implement such a manipulation is in the exposure phase. The exposure phases and memory tests of Experiments 1–2 were designed only to ensure that participants were aware of the clothing and took account of it. There was little to suggest that clothing might be important. And our results were consistent with those of Rácz, Hay, and Pierrehumbert (2020) who also found that "weaker" traits generalized, but not terribly well, in an artificial language task. In Experiment 3 we attempted to change this by directly manipulating the perceived social importance of clothing.

How we interpret variability is also important. In a forced choice task like the one used in Experiments 1–2, participants have to select one of the two options. If they have a *small but consistent* preference for one over the other, the results are likely to look the same as if they have a *strong* preference in the same direction. The difference in variability observed in our results (especially in Experiment 2) implies, therefore, a certain degree of uncertainty and inconsistency in preference rather than the size of the preference per se. We do not know if participants' responses to established aliens in Experiment 2 considered it impossible that a Gili might use the suffix they had seen Nulus use or merely less likely.

In the association test phase of Experiment 3 we therefore also introduced a new task in addition to the forced choice tasks in which participants used a Likert-scale to rate the appropriateness of associations.

4 Experiment 3: the effect of practical social importance

4.1 Experiment overview

As a first step to establishing the components underlying social salience in indexicality, Experiment 3 investigated the effect of contextualizing the role played by clothing. This involved replacing the outfit grouping activity in the familiarization phase with a new activity in which participants were asked to imagine that they were tasked with hosting a diplomatic gathering on the alien planet and had to ensure that wearers of the two different outfits (which for this activity did not covary with species) were equally represented. We manipulated whether participants were told that this was (a) because of the social importance of the outfits or (b) for nonsocial aesthetic reasons. To further examine the degree of uncertainty in the associations established, we also introduced a Likert scale task to the association test phase in which participants were asked to rate the appropriateness of associations.

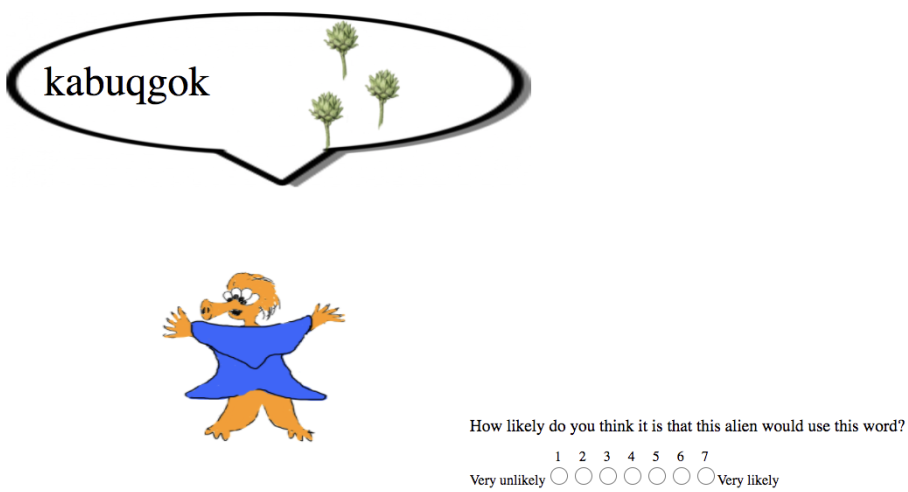


Figure 9: Illustration of Likert scale trials

4.2 Method

4.2.1 Participants

120 participants, recruited through Prolific, took part in return for \$6. After excluding participants ($N = 6$) whose duration was below the 2.5% quantile or above the 97.5% quantile, data from the remaining 114 participants were analyzed (ages 18–56, $median = 22$; 62 women, 50 men, and two who self-identified as "other"). There were 36 participants in the *Baseline condition*, 39 participants in the *Social condition* and 39 participants in the *Nonsocial condition*.

4.2.2 Procedure

Similar to Experiments 1–2, participants in Experiment 3 went through a familiarization phase, a training phase, amemory test, and a final association test phase. However, there were three differences. The first is that, to reduce the overall time of the experiment given other changes, there was now only one training phase and one memory test. (This was in fact consistent with our pilot experiment, reported in Appendix B, and which had rather similar results to Experiment 1.) The second change was that the familiarization phase included a "Diplomatic Gathering" task (replacing the outfit grouping task in the other experiments). The details are discussed in Section 4.2.3. The third change is that, in the association test phase, after the suffix- and alien-selection trials, we introduced a series of Likert-scale trials. In each trial participants saw an image of a single alien wearing one of the two outfits and using a plural word from the language to refer to a familiar referent ($N = 3$ alien species $\times$ 2 outfits $\times$ 8 plural words $=$ 48 trials). Participants were asked to rate on a scale from 1 to 7 how likely they thought it was that the alien would use the word in question (Figure 9).

4.2.3 Diplomatic gathering task and experimental conditions

In the diplomatic gathering task, participants were told that they were hosting a gathering and were responsible for which aliens were invited. They were shown a total number of five screens, each of which contained two delegations of aliens. On every screen they had to select one delegation to invite (Figure 10). Their goal was to ensure that, after they had made all their selections, the two different outfits would be equally distributed across guests. This task was complicated by the fact that the number of aliens wearing each outfit was not necessarily the same on each screen, the competing delegations could be of different sizes, and members of a single delegation did not necessarily wear the same outfit. Furthermore, participants could not be sure of the distribution on upcoming screens. The purpose of this was to make the task sufficiently interesting and challenging that participants would attend to it and, in particular, attend to the outfits. If they succeeded, they were shown a positive message praising their performance. If the outfits were not equally balanced, however, they received negative feedback.

There were two diplomatic gathering conditions (one *Social* and one *Nonsocial* condition), which differed in terms of how the task was framed, as well as a third *Baseline* condition in which there was no diplomatic gathering task and participants simply did the outfit grouping task as in the previous experiments. In the Social condition the importance of balancing outfits for the diplomatic gathering was framed as being a matter of social importance for the aliens. In the Nonsocial condition, by contrast, it was framed as a purely aesthetic concern. The instructions and feedback text can be seen in Appendix C.

4.3 Results for Experiment 3

Mean completion time for Experiment 3 was 30.47 min ($sd = 22$). Mean recall rate in the memory test was 85% ($sd = 9\%$). Figure 11 shows the results of the suffix selection task for both established and new aliens and Figure 13 shows the results of the alien selection task. As in earlier experiments, participants associated linguistic variants with the aliens they had seen use them and the outfits those aliens had worn across the three different conditions, and in both tasks.

Experiment 3 involved more conditions than Experiments 1–2, multiplying the number of comparisons and increasing the risk of a Type 1 error. We therefore configured a mixed-effects logistic regression model using a six-level predictor concatenated with Stimulus (Gili outfit vs. Nulu outfit) and Condition

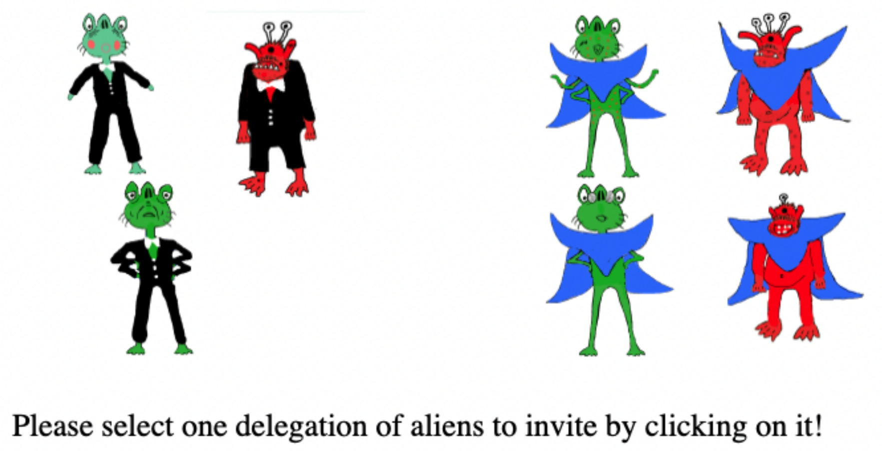


Figure 10: Grouping exercise of Experiment 3

(Baseline vs. Nonsocial vs. Social) as the fixed effect to predict participants' responses. Which version participants were assigned to was included as a version by participant random slope. Word was included as a random intercept. This six-level predictor had six distinct levels: Gili outfit.Baseline, Nulu outfit.Baseline, Gili outfit.Nonsocial, Nulu outfit.Nonsocial, Gili outfit.Social, Nulu outfit.Social. We then extracted post-hoc pairwise comparisons from the fit model (see Table 9 for modeling of new aliens and Table 10 for modeling of established aliens) based on estimated marginal means while correcting for multiple comparisons using Sidak adjustment. Models with Stimulus and Condition and their interaction as fixed effects were also configured (see Table 11). The advantage of our current approach is that it reduced the number of interaction terms to 6 levels of a single fixed effect.

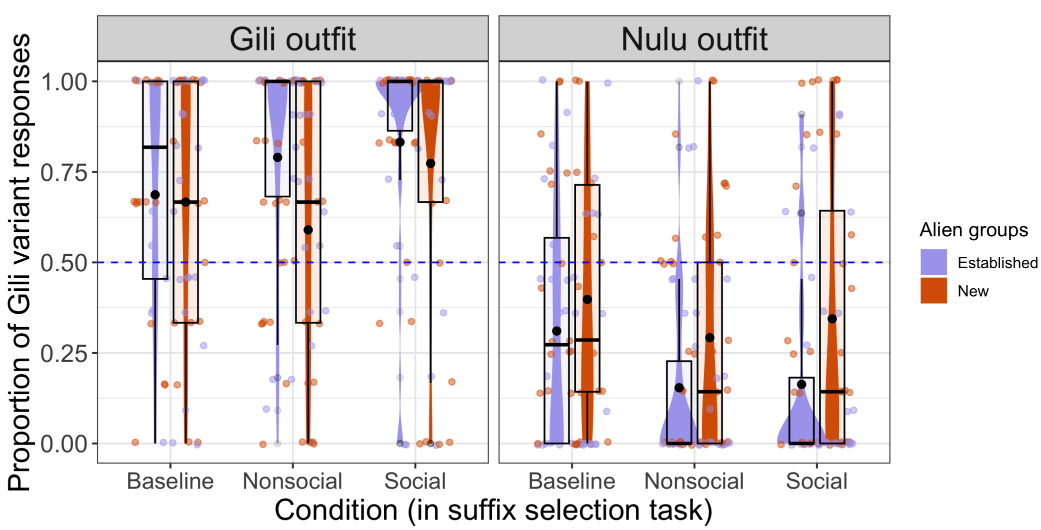


Figure 11: Results for Experiment 3 (Suffix selection): black dots indicate means; purple/orange points indicate participant means; blue dotted line indicates chance level

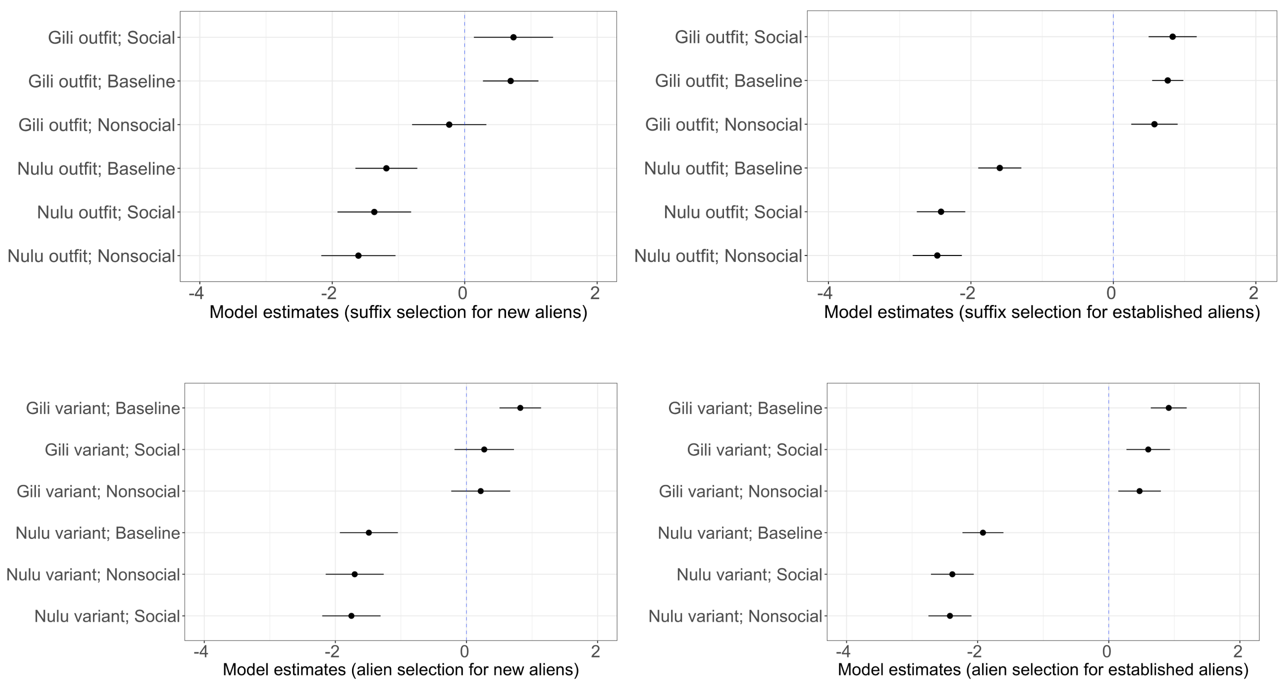


Figure 12: Coefficient for suffix and alien selection in Experiment 3. Error bars are based on 95% confidence intervals.

The results of these pairwise comparisons are displayed in Table 3. These coefficients are further plotted in Figure 12. In all three conditions, in both tasks, and for both established aliens and new aliens, there was a main effect of Stimulus. In other words, participants associated endings with the aliens they had seen use them and the outfits they had seen worn with them. For established aliens, there was also a significant difference between the Nonsocial condition and the Baseline condition ($\beta = 1.45$, $p < 0.001$) and between the Social condition and the Baseline condition ($\beta = 1.65$, $p < 0.001$). There was no difference between the Social and Nonsocial condition, however ($\beta = 0.21$, $p = 0.96$). This pattern of results was true of both tasks. For new aliens, there was no difference between any pair of conditions.

The Likert scale task involved all combinations of alien, outfit, and linguistic variant. An important distinction concerned whether there was a match or a mismatch between what the participants had seen in the Exposure phase. For new aliens (who had not been seen in the Exposure phase) a mismatch could mean only that the linguistic variant did not match the outfit worn. For established aliens, there could be a mismatch between the linguistic variant and either the outfit or the alien (or both). As can be seen in Figures 14 and 15, the general pattern was that—as expected—participants rated matches higher than mismatches. However, this pattern appeared to be considerably more pronounced for (mis)matches between linguistic variants and established aliens than for (mis)matches between linguistic variants and outfits, whether for established or new aliens. To test these differences statistically, we used the same approach as for the main results presented above (see model details in Table 12 and Table 13). A full set of comparisons can be seen in Table 4, which is further illustrated visually through Figure 16. There was a significant difference between matched and mismatched trials for all social and Nonsocial conditions. There was also a significant difference in the Baseline condition for (mis)matches between linguistic variant and established alien species and for outfits worn by new aliens. However, there was no significant

Table 3: Post-hoc comparisons for suffix selection and alien selection in Experiment 3

Suffix selection: established aliens	Estimate	Std. Error	z value	Pr(>|z|)
Baseline (Gili outfit - Nulu outfit)	1.58	0.15	10.32	<0.001
Nonsocial (Gili outfit - Nulu outfit)	3.02	0.18	16.95	<0.001
Social (Gili outfit - Nulu outfit)	3.24	0.18	17.61	<0.001
Nonsocial (Gili outfit - Nulu outfit) - Baseline (Gili outfit-Nulu outfit)	1.45	0.24	6.15	<0.001
Social (Gili outfit - Nulu outfit) - Baseline (Gili outfit-Nulu outfit)	1.65	0.24	6.91	<0.001
Social (Gili outfit - Nulu outfit) - Nonsocial (Gili outfit-Nulu outfit)	0.21	0.26	0.80	0.96
Suffix selection: new alien	Estimate	Std. Error	z value	Pr(>|z|)
Baseline (Gili outfit - Nulu outfit)	1.18	0.24	4.96	<0.001
Nonsocial (Gili outfit - Nulu outfit)	1.37	0.23	5.89	<0.001
Social (Gili outfit - Nulu outfit)	2.10	0.26	8.25	<0.001
Nonsocial (Gili outfit - Nulu outfit) - Baseline (Gili outfit-Nulu outfit)	0.19	0.33	0.57	1
Social (Gili outfit - Nulu outfit) - Baseline (Gili outfit-Nulu outfit)	0.92	0.35	2.66	0.05
Social (Gili outfit - Nulu outfit) - Nonsocial (Gili outfit-Nulu outfit)	0.73	0.34	2.14	0.18
Alien selection: established aliens	Estimate	Std. Error	z value	Pr(>|z|)
Baseline (Gili variant - Nulu variant)	1.91	0.16	12.05	<0.001
Nonsocial (Gili variant - Nulu variant)	2.89	0.17	16.59	<0.001
Social (Gili variant - Nulu variant)	2.99	0.18	16.88	<0.001
Nonsocial (Gili variant - Nulu variant) - Baseline (Gili variant - Nulu variant)	0.97	0.24	4.13	<0.001
Social (Gili variant - Nulu variant) - Baseline (Gili variant - Nulu variant)	1.07	0.24	4.49	<0.001
Social (Gili variant- Nulu variant) - Nonsocial (Gili variant -Nulu variant)	0.10	0.25	0.39	1
Alien selection: new aliens	Estimate	Std. Error	z value	Pr(>|z|)
Baseline (Gili variant - Nulu variant)	1.49	0.23	6.60	<0.001
Nonsocial (Gili variant - Nulu variant)	1.92	0.23	8.48	<0.001
Social (Gili variant - Nulu variant)	2.03	0.23	8.84	<0.001
Nonsocial (Gili variant - Nulu variant) - Baseline (Gili variant -Nulu variant)	0.43	0.32	1.35	0.69
Social (Gili variant - Nulu variant) - Baseline (Gili variant - Nulu variant)	0.54	0.32	1.67	0.45
Social (Gili variant - Nulu variant) - Nonsocial (Gili variant - Nulu variant)	0.10	0.32	0.32	1

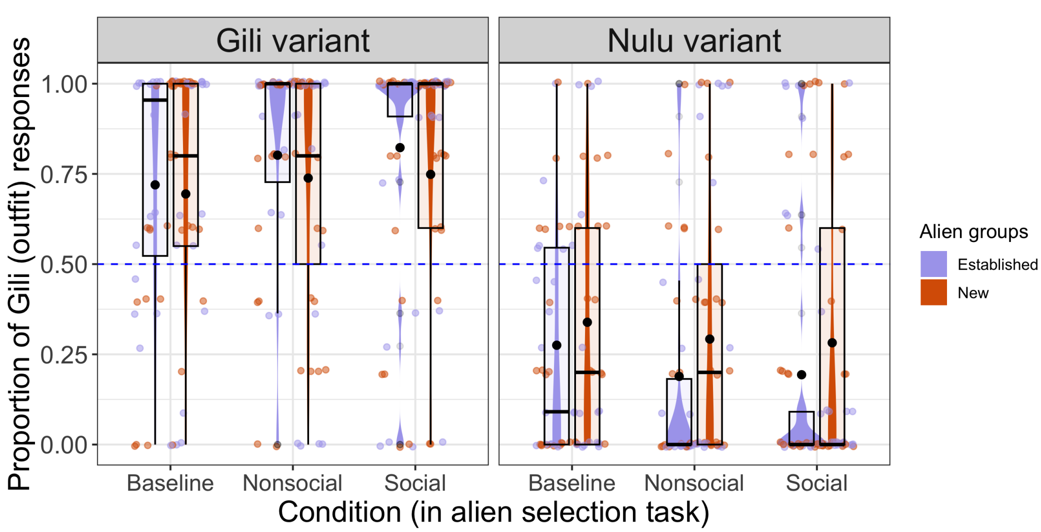
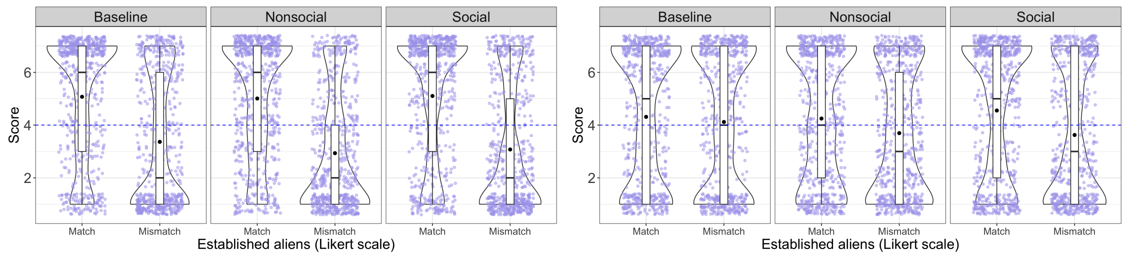


Figure 13: Results for Experiment 3 (Alien selection): black dots indicate means; purple/orange points indicate participant means; blue dotted line indicates chance level

difference for outfits worn by established aliens, suggesting that associations formed with alien species tended to mask the role of outfits. There was, nonetheless, a significant difference between the Social condition and the Baseline condition, suggesting that emphasizing the social importance of the outfits might have strengthened the association for participants, consistent with the effects reported for the alien-selection and suffix-selection tasks. Nonetheless, a similar pattern was not seen for new aliens. For new aliens, there was instead a significant difference between the Nonsocial and the Baseline and not between the Social and the Baseline. In general, the effects of the manipulation in this experiment seem to have been subtle and the results should be interpreted with caution.



(a) (b)

Figure 14: Likert scale results for established aliens in Experiment 3 (left: alien-variant match; right: outfit-variant match)

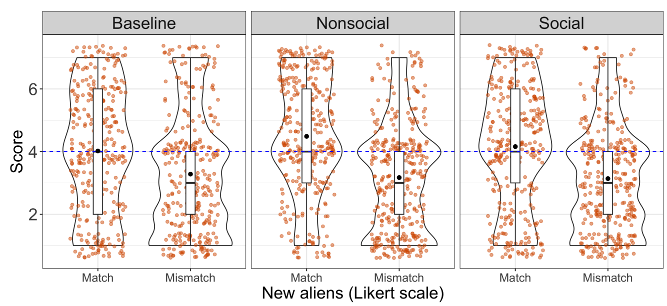


Figure 15: Likert-scale results for new aliens in Experiment 3 (outfit-variant match)

4.4 Discussion of Experiment 3

Experiment 3, as a step towards better understanding the role played by salience (and what salience means) in establishing indexicality, tested how associations between speaker groups, outfits and linguistic forms would be modulated by the way outfits were contextualized. Our results first established that the role of outfits mattered in how participants perceived the associations between speaker groups, clothing and plural endings. When we emphasized the social or aesthetic significance of clothing, this strengthened the linguistic associations participants formed with it. However, we did not see a clear difference between the Nonsocial and Social conditions, suggesting that attaching importance to a trait may matter more for indexicality than the kind of importance attached. We should be cautious of how we interpret results with regard to this question, however. Given that aesthetic questions can have their own social consequences, and both conditions (for the sake of control) invoked the idea of an important social event,

Table 4: Post-hoc comparisons for likert scale in Experiment 3

Likert scale: new alien (outfit-variant match vs. outfit-variant mismatch)	Estimate	Std. Error	z value	Pr(>|z|)
Baseline (Match - Mismatch)	0.38	0.07	5.35	<0.001
Nonsocial (Match - Mismatch)	0.69	0.07	9.97	<0.001
Social (Match - Mismatch)	0.53	0.07	7.76	<0.001
Nonsocial (Match - Mismatch) - Baseline (Match - Mismatch)	0.30	0.10	3.04	0.01
Social (Match - Mismatch) - Baseline (Match - Mismatch)	0.15	0.10	1.51	0.57
Social (Match - Mismatch) - Nonsocial(Match - Mismatch)	-0.15	0.10	-1.57	0.52
Likert scale: established aliens (alien-variant match vs. alien-variant mismatch)	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Baseline (Match - Mismatch)	0.69	0.05	13.11	<0.001
Nonsocial (Match - Mismatch)	0.83	0.05	16.74	<0.001
Social (Match - Mismatch)	0.81	0.05	16.25	<0.001
Nonsocial (Match - Mismatch) - Baseline (Match - Mismatch)	0.15	0.07	2.02	0.24
Social (Match - Mismatch) - Baseline (Match - Mismatch)	0.13	0.07	1.76	0.39
Social (Match - Mismatch) - Nonsocial(Match - Mismatch)	-0.02	0.07	-0.26	1
Likert scale: established aliens (outfit-variant match vs. outfit-variant mismatch)	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Baseline (Match - Mismatch)	0.08	0.06	1.46	0.61
Nonsocial (Match - Mismatch)	0.22	0.05	4.14	<0.001
Social (Match - Mismatch)	0.37	0.05	6.91	<0.001
Nonsocial (Match - Mismatch) - Baseline (Match - Mismatch)	0.14	0.08	1.80	0.36
Social (Match - Mismatch) - Baseline (Match - Mismatch)	0.29	0.08	3.72	<0.01
Social (Match - Mismatch) - Nonsocial(Match - Mismatch)	0.15	0.08	1.99	0.25

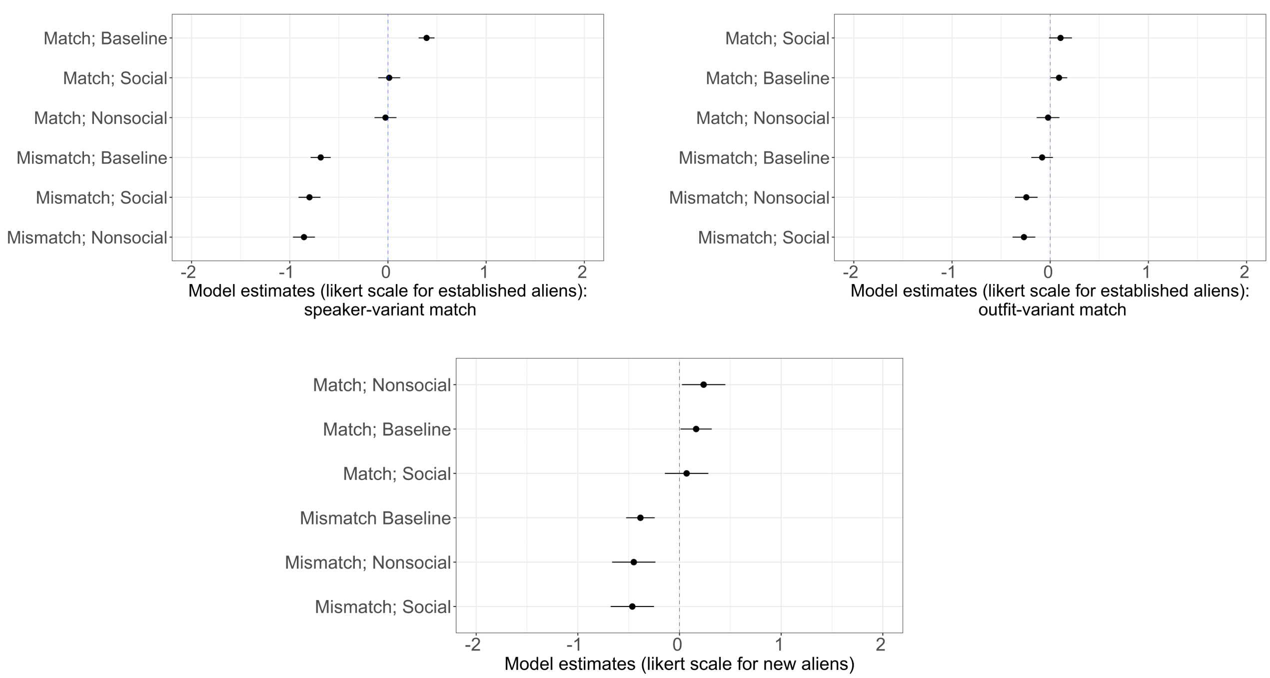


Figure 16: Coefficient for likert scale in Experiment 3

it is possible that participants also interpreted the aesthetic value of outfits in social terms. The strength of the Social condition might also have been reduced by the non-specificity of the social significance. In real life, clothing-based indexicality is typically connected with some real social context. In future work this could be explored by invoking particularly important social categories, such as religion.

For the Likert-scale trials, furthermore, the effects of the manipulation seem to have been rather subtle. In future work it would be interesting to investigate this question in different ways, including more clearly non-social contexts, or by drawing attention to outfits in a negative way to control for the role of increased attention (cf. Roberts & Fedzechkina, 2018). Along similar lines, it is worth noting that the experiment itself was a learning task without an interactive component. It is possible that making the tasks themselves more practically social would influence the relative importance of social and nonsocial value attached to traits (cf. Sneller & Roberts, 2018, and Wade & Roberts, 2020, who employed such an interactive task in related work, and Fedzechkina, Hall Hartley, & Roberts, 2022, who did not and voiced a similar concern.)

5 General Discussion

We investigated the emergence of different orders of sociolinguistic indexicality in a laboratory setting using an artificial-language-learning paradigm. Experiment 1 investigated whether participants exposed to linguistic variation that covaried reliably with both a group of speakers (the alien species) and a nonlinguistic cultural trait (their clothing) would form an association and, if so, whether they would associate the linguistic variation primarily with the speaker group or with both the speaker group and the nonlinguistic cultural trait. We found that participants tended to overwhelmingly associate the variable plural endings with alien species rather than their outfits. In Experiment 2, we introduced a new alien species exhibiting the same variation in cultural traits (i.e., wearing the same outfits) to investigate whether participants would extend established associations to the new aliens via clothing. We found that this occurred. Participants again strongly associated suffixes with established aliens but now extended associations to new aliens via clothing, indicating that, in spite of the strong primary association they had made with species, they had also formed a secondary association with clothing. In Experiment 3, we tested the hypothesis that associations would be strengthened by attaching social or aesthetic significance to the outfits, and we found that they were. We did not, however, find strong evidence that it mattered whether this significance was of a social or aesthetic nature.

In summary, our results suggest that linguistic indexicality can arise through exposure to co-occurring linguistic and nonlinguistic variables and the extension of established associations to new contexts in which the indexed trait is dissociated from the originally observed bearers. Indexicality is modulated by the social salience and practical importance of the nonlinguistic variables in question.

The role of co-occurrence in the formation of indexical associations is perhaps worth putting into a broader context within cognitive sociolinguistics. A number of studies have, for instance, investigated the role of co-occurring non-linguistic cues—such as conversation topics or even stuffed toys—in influencing sociolinguistic perception (e.g., Love & Walker, 2013; Hurring et al., 2022; see Campbell-Kibler, 2010 for further discussion of such questions in sociolinguistic cognition), while other work has focused on the role of co-occurring linguistic cues in motivating *expectation-based* conversational convergence on linguistic variants that are not themselves observed (e.g., Wade, 2022; see also Wade & Roberts, 2020 for an artificial-language approach to this question). This work points to the role of top-down cues in language processing (and production) given established sociolinguistic categories. One way of thinking about our own study is to treat the emergence of indexicality as a process of identifying useful top-down cues for future processing.

In this context, something we might call salience is likely to play an important part, and is likely to be important to our participants' tendency to associate linguistic variants primarily with speaker groups and only secondarily with their clothing. The term salience is used rather widely in the language sciences, and it has been noted that it is not straightforward to pin down precisely what is meant by it (Jaeger & Weatherholtz, 2016; Rácz, 2013; Boswijk & Coler, 2020), although all interpretations seem to share the notion of salience as a property of some unit that renders it perceptually more prominent in an array of competing units (Rácz, 2013; Wolfe & Horowitz, 2004). More specifically, Wolfe and Horowitz (2004) argued that sensory information provides an "embarrassment of riches" that the brain is unable to process all at once, requiring attention to be preferentially given to certain objects in (in that case) the visual field. On the whole, understanding of the properties of salience seems more developed in the field of visual perception (Treue, 2003; Rust & Cohen, 2022), and it is in fact possible that the effect in our experiment was driven primarily by visual salience. However, we should be at least somewhat wary of this interpretation, at least on a simplistic bottom-up interpretation of what visual salience means. First, we made an effort to make the outfits highly distinctive and to reinforce attention to them using the familiarization and memory tasks. Second, indexicality also seems to attach preferentially to speaker group in real-world sociolinguistic behavior (Silverstein, 2003; Eckert, 2008), which seems even less likely to be consistently driven by *visual* salience.

We therefore consider it more probable that, while visual salience likely played some role, our result in this respect was driven by the fact that speaker groups are usually more socially salient than clothing more generally, particularly when that clothing (as in Experiments 1–2) has no clear social meaning attached to it. This is also consistent with the findings of Rácz et al. (2017), whose participants were more successful in learning an association with a socially salient feature such as gender than with a socially non-salient feature such as spatial orientation. Our participants' bias towards species as opposed to clothing is consistent for that matter with the role of top-down social cues in visual salience (see, for instance, work by Buschman & Miller, 2007, and Saalmann, Pigarev, & Vidyasagar, 2007, on the role of top-down processes in directing visual attention selectively at entities that are likely to be of particular importance). That is, we consider that a useful way to interpret our results might be to say that, in processing the context of the variation in the alien language, our participants likely reduced the complexity of the task by focusing attention, out of the available contextual factors, on the one that seemed most likely to be socially important. More broadly, this would suggest that orders of indexicality might essentially track orders of social salience. That is, the establishment of a new order of indexicality would—in this interpretation—involve the selection of whatever is the most socially salient, or relevant subtrait of whatever category is involved in the lower-order indexicality. It is nonetheless a limitation of our current study that we did not measure degrees of social salience in clothing versus species. This would be a very important focus for future work as it would allow us to identify with considerably more precision and detail what is really meant by social salience or social relevance and the role it plays in enregisterment and indexicality.

Only in Experiment 3 of the present study did we explicitly manipulate social salience, and we found some evidence that this does indeed modulate the strength of indexical associations. However, such results did not show up robustly across tasks and conditions. There were interesting differences in particular between results for new aliens and those for established aliens. A likely explanation for this is that it was motivated by the relative strength of associations that participants had formed. For new aliens, the association was entirely mediated by outfit, and in all conditions this was a weaker, secondary association compared with the associations formed with the aliens seen in the Exposure phase. As suggested above, in the Discussion for Experiment 3, our manipulation was in some respects a relatively subtle one, and the social meaning invoked by it was non-specific. There were no *real* consequences (such as loss of monetary reward) for failure in the Diplomatic gathering task. Nor is it fully guaranteed that we succeeded in eliminating social interpretations of the nonsocial version of the task. This should be remedied in future work. A task incorporating real interaction (cf. Wade & Roberts, 2020; Sneller & Roberts, 2018) or simulated interaction (cf. Buz, Tanenhaus, & Jaeger, 2016) might also help to bolster the social context; monetary rewards or penalties could also be employed, such as by having participants gain and lose points during the experiment with the total corresponding to a bonus payment after participating.

Along with the differences between conditions, we also saw differences between tasks. The first important difference concerns the alien-selection and the suffix-selection tasks. Broadly speaking, the alien-selection task (in which participants were asked which alien might use a given form) tended to produce starker results than suffix selection (in which participants were asked which form a given alien might use). That is, variability was lower for alien selection than suffix selection; this was particularly obvious in Experiment 1 (Figure 5). The reverse pattern did not occur in any condition. This could be due to chance, although it mirrors Lai et al.'s (2020) results, who found much the same thing—more variability in one selection task than in its converse—with a less deterministic relationship between linguistic and nonlinguistic traits. There, the authors made an analogy with natural language variation. If asked who pronounces "about" as "aboot", for instance, people familiar with relevant stereotypes are likely to reliably mention Canadians, even though so-called Canadian raising is also a feature of many US varieties and some Canadians do not exhibit it. If asked how a Canadian might pronounce the word "about", on the other hand, there seems likely to be greater variability in responses, even if the majority of respondents respond with a raised variant. "Aboot", in other words, fairly straightforwardly indexes Canadians, but the reverse is not so obviously the case. Indexicality is not necessarily a symmetrical two-way association, even—as in our experiment—if the relationship it is based on is reliably one-to-one. (See Campbell-Kibler, 2007, for a discussion of related asymmetries.)

The other task difference came in Experiment 3 with the introduction of the Likert-scale task. The pattern of note here is that, for established aliens, results looked broadly similar to results in the forced-choice tasks. For new aliens, however, the Likert-scale results implied much less certainty about associations, consistent with the greater variability evident in the forced-choice results for new aliens. The point is that a relatively clear difference in association revealed in one kind of task need not correspond to a strongly dichotomous perception of how the relevant relationships operate. Taken together, these different task effects are a reminder of the importance of how one asks a question to what kind of answer is received.

In the discussion above we have brought up several ways in which future research might build on what we have presented here. In particular we would emphasize the utility of experimental paradigms that involve interaction with other individuals. For questions of social meaning in which it is important to generate investment in that meaning, this would appear to be a particularly fruitful approach (Sneller & Roberts, 2018; Roberts & Sneller, 2020). Another potential limitation in the current study concerns the choice of clothing in particular as a cultural trait. This had the advantage of being easy to make salient in a simple way and being easily transferable between individuals for the purposes of our manipulations; but clothing has the disadvantage of very often being itself an arbitrary index of other variables rather than being a primary variable (such as a personality trait) in its own right. Again, more options become available in more interactive paradigms. For instance, Sneller and Roberts (2018), varied aliens' toughness in an interactive game in which participants could fight each other for resources, while Wade and Roberts (2020) manipulated perceived and expected interlocutor identity. Another limitation concerns the deterministic one-to-one mapping between clothing, linguistic variants, and alien species. This was done to simplify the task and design, given that there was a three-way relationship (and, in Experiments 2 and 3, three different alien species). However, it does not reflect patterns of linguistic variation in the real-world, which have long been known to exhibit orderly heterogeneity (Weinreich et al., 1968). In future work, it would be important to incorporate probabilistic variation (cf. Lai et al., 2020).

Here we have presented, to our knowledge, the first laboratory study focusing specifically on the emergence of sociolinguistic indexicality, of different kinds, through exposure to covariation between linguistic variants and nonlinguistic traits and extension to novel contexts. We hope that this will be built upon in future work.

References

Acerbi, A., Ghirlanda, S., & Enquist, M. (2012). The logic of fashion cycles. *PLOS One*, *7*(3), e32541.

Agha, A. (2003). The social life of cultural value. *Language & communication*, *23*(3-4), 231–273.

Agha, A. (2007). *Language and social relations* (Vol. 24). Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (Second ed.). Cambridge University Press.

Blank, A. (1999). Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. *Cognitive Linguistics Research*, 61–89.

Borgida, E., & Howard-Pitney, B. (1983). Personal involvement and the robustness of perceptual salience effects. *Journal of Personality and Social Psychology*, *45*(3), 560.

Boswijk, V., & Coler, M. (2020). What is salience? *Open Linguistics*, *6*(1), 713–722.

Burridge, K. (2012). Euphemism and language change: The sixth and seventh ages. *Lexis. Journal in English Lexicology*(7).

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*(5820), 1860–1862.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68–86.

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

Campbell-Kibler, K. (2010). New directions in sociolinguistic cognition. *University of Pennsylvania Working Papers in Linguistics*, *15*(2), 5.

Dahan, D., & Brent, M. R. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, *128*(2), 165.

Denton, S. E., Kruschke, J. K., & Erickson, M. A. (2008). Rule-based extrapolation: A continuing challenge for exemplar models. *Psychonomic bulletin & review*, *15*(4), 780–786.

Dyer, J. (2007). Language and identity. In C. Llamas, L. Mullany, & P. Stockwell (Eds.), (pp. 101–108). Routledge.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476.

Fedzechkina, M., Hall Hartley, L., & Roberts, G. (2022). Social biases can lead to less communicatively efficient languages. *Language Acquisition*, 1–26.

Fischer, L. M., Meyers, C., Cummins, R. G., Gibson, C., & Baker, M. (2020). Creating relevancy in agricultural science information: Examining the impact of motivational salience, involvement and pre-existing attitudes on visual attention to scientific information. *Journal of Applied Communications*, *104*(2), 1051–0834.

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

Guy, G. R., & Hinskens, F. (2016). Linguistic coherence: Systems, repertoires and speech communities. *Lingua*, *172*(173), 1–9.

Havas, V., Waris, O., Vaquero, L., Rodríguez-Fornells, A., & Laine, M. (2015). Morphological learning in a novel language: A cross-language comparison. *Quarterly Journal of Experimental Psychology*, *68*(7), 1426–1441.

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*(2), 151–195.

Hughes, G. (1990). What is register? *English Today*, *6*(2), 47–51.

Hurring, G., Hay, J., Drager, K., Podlubny, R., Manhire, L., & Ellis, A. (2022). Social priming in speech perception: Revisiting kangaroo/kiwi priming in New Zealand English. *Brain Sciences*, *12*(6), 684.

Irvine, J. (2001). "style" as distinctiveness: The culture and ideologies of linguistic differentiation. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 21–41). Cambridge: Cambridge University Press.

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2013). What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, *36*(7), 1469–1482.

Jaeger, T. F., & Weatherholtz, K. (2016). What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology*, *7*, 1115.

Jaffe, A. (Ed.). (2009). *Stance: Sociolinguistic perspectives*. Oxford: Oxford University Press.

Johnstone, B. (2016). Enregisterment: How linguistic items become linked with ways of speaking. *Language and Linguistics Compass*, *10*(11), 632–643.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81.

Kwon, M.-K., Setoodehnia, M., Baek, J., Luck, S. J., & Oakes, L. M. (2016). The development of visual search in infancy: Attention to faces versus salience. *Developmental psychology*, *52*(4), 537.

Lai, W., Rácz, P., & Roberts, G. (2020). Experience with a linguistic variant affects the acquisition of its sociolinguistic meaning: An alien-language-learning experiment. *Cognitive science*, *44*(4), e12832.

Lenth, R., & Lenth, M. R. (2018). Package 'lsmeans'. *The American Statistician*, *34*(4), 216–221.

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

Llamas, C., Watt, D., & MacFarlane, A. E. (2016). Estimating the relative sociolinguistic salience of segmental variables in a dialect boundary zone. *Frontiers in Psychology*, *7*, 1163.

Love, J., & Walker, A. (2013). Football versus football: Effect of topic on /r/ realization in American and English sports fans. *Language and Speech*, *56*(4), 443–460.

Ma, T., & Komarova, N. L. (2019). Object-label-order effect when learning from an inconsistent source. *Cognitive Science*, *43*(8), e12737.

Merkx, M., Rastle, K., & Davis, M. H. (2011). The acquisition of morphological knowledge investigated through artificial language learning. *Quarterly Journal of Experimental Psychology*, *64*(6), 1200–1220.

Meyerhoff, M., & Schleef, E. (2012). Variation, contact and social indexicality in the acquisition of (ing) by teenage migrants. *Journal of Sociolinguistics*, *16*(3), 398–416.

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

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

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*, 1–31.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, *11*(4), 478–504.
R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Rácz, P. (2013). *Salience in sociolinguistics: A quantitative approach*. De Gruyter Mouton.
Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in psychology*, *8*, 51.
Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2020). Not all indexical cues are equal: Differential sensitivity to dimensions of indexical meaning in an artificial language. *Language Learning*, *70*(3), 848–885.
Roberts, G. (2017). The linguist's *Drosophila*: Experiments in language change. *Linguistics Vanguard*, *3*(1).
Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171*, 194–201.
Roberts, G., & Sneller, B. (2020). Empirical foundations for an integrated study of language evolution. *Language Dynamics and Change*, *10*(2), 188–229.
Rust, N. C., & Cohen, M. R. (2022). Priority coding in the visual system. *Nature Reviews Neuroscience*, *23*(6), 376–388.
Saalmann, Y. B., Pigarev, I. N., & Vidyasagar, T. R. (2007). Neural mechanisms of visual attention: How top-down feedback highlights relevant locations. *Science*, *316*(5831), 1612–1615.
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.
Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & communication*, *23*(3-4), 193–229.
Smith, J. J. (2013). *Essentials of Early English: Old, Middle and Early Modern English*. Routledge.
Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.
Sneller, B., & Roberts, G. (2018). Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition*, *170*, 298–311.
Treue, S. (2003). Visual attention: The where, what, how and why of saliency. *Current opinion in neurobiology*, *13*(4), 428–432.
Vihman, V.-A., Nelson, D., & Kirby, S. (2018). Animacy distinctions arise from iterated learning. *Open Linguistics*, *4*(1), 552–565.
Wade, L. (2022). Experimental evidence for expectation-driven linguistic convergence. *Language*.
Wade, L., & Roberts, G. (2020). Linguistic convergence to observed versus expected behavior in an alien-language map task. *Cognitive Science*, *44*(4), e12829.
Wang, J., Chandler, D. M., & Le Callet, P. (2010). Quantifying the relationship between visual salience and visual importance. In *Human vision and electronic imaging xv* (Vol. 7527, pp. 160–168).
Weinreich, U., Labov, W., & Herzog, M. (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics* (pp. 95–195). Austin: University of Texas Press.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.
Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, *5*(6), 495–501.

A Object stimuli images used in Experiment 1–3

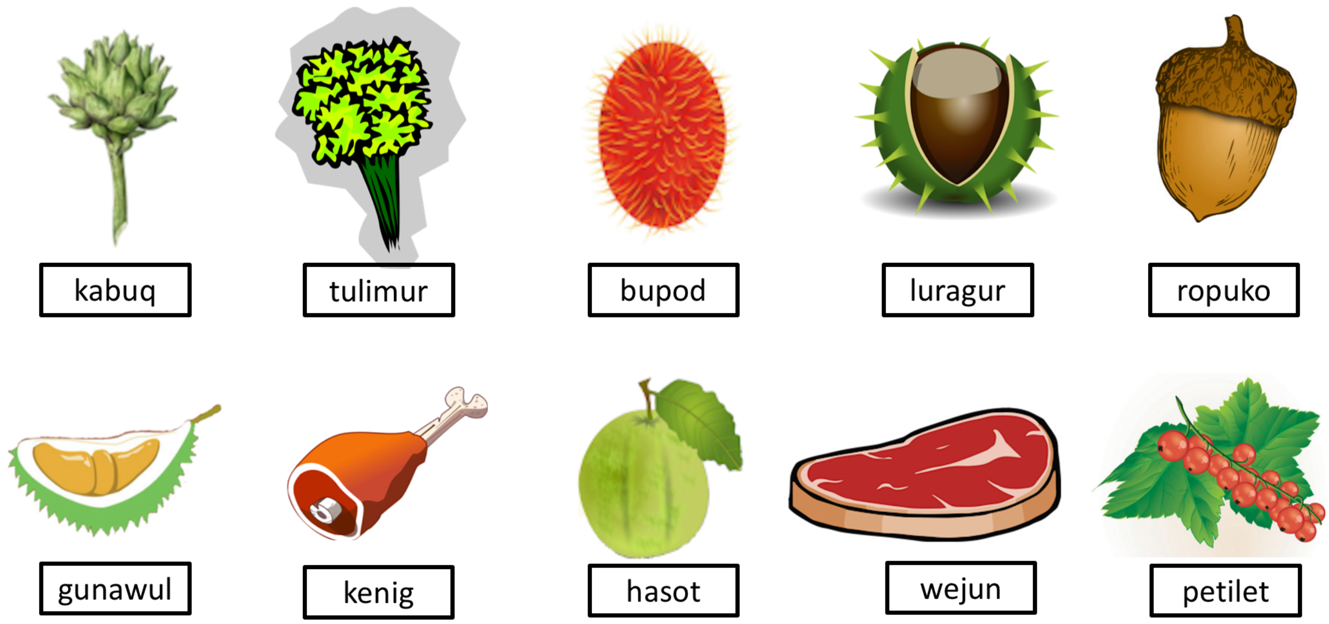
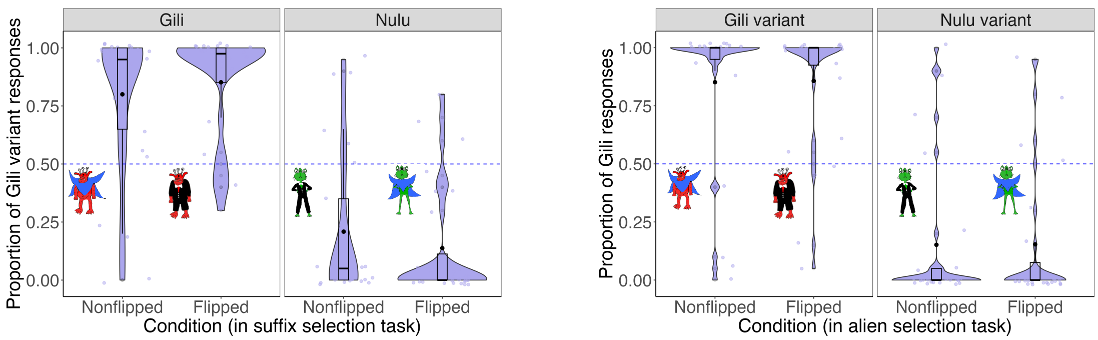


Figure 17: Object stimuli images used in the current study

B Initial pilot

Before conducting Experiment 1, we conducted an initial version of that experiment, which was identical in its design to Experiment 1 except that participants were exposed to only one training phase and one memory test phase. For this experiment there was also a software error that meant that aliens and outfits were not counterbalanced as they should have been. We include the results of this experiment, which we will refer to as Experiment 1a, here for the purpose of comparison.

Mean completion time was 22.81 min (sd = 12). Outliers were excluded as described in Section 2.2.1. The mean success rate on the 12 memory trials was 86.1% (sd = 6.6%). Figure 18 shows the aggregate results of suffix selection and alien selection in Experiment 1a. In both the suffix- and the alien-selection tasks, participants strongly associated suffixes with alien species. There was, however, more variability in the suffix-selection task than in the alien selection task (sds = 0.40, 0.36 respectively).



(a) (b)

Figure 18: Results for Experiment 1a (black dots indicate means; purple dots indicate outliers based on 95% CI; blue dotted line indicates chance level): (a): suffix selection task; (b): alien selection task

Mixed-effects logistic regression models were fit separately for the two tasks, with Response (suffix selection: Gili variant; Alien selection: Gili) as the dependent variable, Condition (Nonflipped as the reference level), Stimuli (Gili and the Gili Outfit as the reference level in suffix selection, and the Gili suffix as the intercept in alien selection) and their interaction as independent variables, and Participant and Word as random intercepts. by-Participant for Stimuli was included as a random slope. Model results revealed that for the SUFFIX SELECTION, there was a significant effect of Stimulus ($\beta = -6.29$, $p < 0.001$, for alien stimulus and $\beta = -2.72$, $p < 0.001$, for outfit stimulus). That is, participants tended to match suffixes with the aliens and outfits they had seen associated with them in the training phase. There was also a significant main effect of Condition, but only for outfit stimulus ($\beta = -3.22$, $p < 0.001$), not for alien stimulus ($\beta = 0.84$, $p = 0.36$). This implied that participants' tendency to match suffixes with the aliens and outfits they had seen during training became significantly stronger in the Flipped condition for outfit stimulus, when aliens were in unexpected outfits. There was also a significant interaction between Stimulus and Condition for outfit stimulus ($\beta = 6.30$, $p < 0.001$) but not for alien stimulus ($\beta = -2.04$, $p = 0.28$). Post-hoc pairwise comparisons were further implemented using estimated marginal means. The results further suggested that overall, the response difference in response to different aliens was not significantly different between the Flipped condition and the Nonflipped condition ($\beta = 0.84$, $p = 0.31$). For Alien selection, there was a significant effect of (suffix) Stimulus ($\beta = -3.47$, $p < 0.001$, for alien responses and $\beta = -3.47$, $p < 0.001$, for outfit responses), suggesting that participants were significantly less likely to select a Gili alien than a Nulualien when presented with a Gili suffix. The Condition effect was found only for outfit responses2 ($\beta = -3.54$, $p < 0.001$), not for alien responses ($\beta = 0.04$, $p = 0.80$, n.s.). This suggests that, at least for outfit responses, when Nulus wore a Gili outfit, participants were significantly less likely to choose the Gili variant, as opposed to the Nulu variant. The interaction between Stimulus and Condition was significant only for outfit responses ($\beta = 6.97$, $p < 0.001$) but not for alien responses ($\beta = -0.03$, $p = 0.90$), n.s.). Post-hoc pairwise comparisons based on estimated marginal means further suggested again that the Flipped condition was not significantly different from the Nonflipped condition ($\beta = 1.03$, $p = 0.81$).

C Instructions and feedback for Experiment 3

Social condition

Instructions

Now that you have familiarized yourself with the aliens on Ginuli, it's time to make arrangements for your first event. You will be hosting a diplomatic gathering and inviting different delegations of aliens to it. It is very important that we make a good impression here. One way to do this involves paying attention to the aliens' outfits. Clothing matters a lot to the inhabitants of Ginuli, who care a lot about who wears what outfit. As you will have noticed, some aliens on Ginuli wear blue outfits and some wear black outfits, and each of these two groups (the aliens who wear blue outfits and the ones who wear black outfits) is likely to be annoyed if they think the other group is getting preferential treatment. We should therefore try to invite an equal number of blue-outfitted and black-outfitted aliens to the gathering. On each of the next few screens you will be presented with two delegations of aliens (often of different sizes) and you must pick one in each case. In doing so, please try to keep an equal number of blue- and black-outfitted aliens overall. If we end up offending one of the groups, we will not make the impact we hope for!

Feedback

- In case of failure: "Oh no! There were more guests wearing one outfit than the other! This caused offense! Let's hope we can do better from now on!".
- In case of success: "Congratulations! You kept the blue-outfitted aliens evenly balanced with the black-outfitted aliens, so no one was offended and the party was a success! Let's hope we continue to do so well!"

Nonsocial condition

Instructions

Now that you have familiarized yourself with the aliens on Ginuli, it's time to make arrangements for your first event. You will be hosting a diplomatic gathering and inviting different delegations of aliens to it. It is very important that we make a good impression here. One way to do this involves paying attention to the aliens' outfits. Clothing aesthetics matter a lot to the inhabitants of Ginuli, who care a lot about how their outfits look. As you will have noticed, aliens on Ginuli sometimes wear blue outfits and sometimes wear black outfits. The aliens really like balanced aesthetics, and they are likely to be annoyed if the colors aren't properly balanced. We should therefore try to invite an equal number of blue-outfitted and black-outfitted aliens to the gathering. On each of the next few screens you will be presented with two delegations of aliens (often of different sizes) and you must pick one in each case. In doing so, please try to keep an equal number of blue- and black-outfitted aliens overall. If the colors end up unbalanced, we will not make the impact we hope for!

Feedback messages

- In case of failure: “Oh no! There were more guests wearing one outfit than the other! The guests weren’t happy! Let’s hope we can do better from now on!”.
- In case of success: “Congratulations! You kept the outfit colors evenly balanced, so our guests were happy and the party was a success! Let’s hope we continue to do so well!”.

D Models in Experiment 1–3

Experiment 1

Table 5: Model results for Experiment 1: Response $\sim$ Stimuli * Condition + (1+Stimuli|Participant) + (1|Word)

Task	Fixed Effects	Estimate	SE	z value	Pr(>|z|)
Suffix selection: Predict Gili variant responses by Condition and Stimuli (alien species)	Intercept	5.76	1.04	5.55	<0.001
	Stimuli (Nulu)	-10.62	1.81	-5.86	<0.001
	Condition (Flipped)	-2.28	1.19	-1.92	0.05
	Stimuli (Nulu alien) : Condition (Flipped)	4.25	2.15	1.98	0.05
Suffix selection: Predict Gill variant responses by Condition and Stimuli (alien outfit)	Intercept	5.45	0.95	5.73	<0.001
	Stimuli (Nulu outfit)	-10.52	1.79	-5.89	<0.001
	Condition (Flipped)	-8.42	1.28	-6.59	<0.001
	Stimuli (Nulu outfit): Condition (Flipped)	16.82	2.40	7.00	<0.001
Alien selection: Predict Gili responses by Condition and Stimuli (suffix)	Intercept	9.58	1.52	6.31	<0.001
	Stimuli (Nulu suffix)	-21.03	3.25	-6.48	<0.001
	Condition (Flipped)	-0.14	1.55	-0.09	0.93
	Stimuli (Nulu suffix): Condition (Flipped)	1.06	3.38	0.31	0.75
Alien selection: Predict Gili outfit responses by Condition and Stimuli (suffix)	Intercept	9.57	1.50	6.40	<0.001
	Stimuli (Nulu suffix)	-21.00	3.19	-6.58	<0.001
	Condition (Flipped)	-19.00	2.53	-7.50	<0.001
	Stimuli (Nulu suffix): Condition (Flipped)	40.95	5.18	7.90	<0.001

Table 6: Post-hoc pairwise test results of the model fit differences between Stimuli at Condition level for Experiment 1 (with Tukey's adjustment): suffix selection task

Contrast	Pair	Estimate	SE	z.ratio	p.value
Nonflipped	Gili - Nulu	40729.49	73629.88	5.87	<0.001
Flipped	Gili - Nulu	580.12	820.76	4.50	<0.001
Gili	Flipped - Nonflipped	0.10	0.12	-1.93	0.22
Nulu	Flipped - Nonflipped	7.18	7.39	1.92	0.22

Table 7: Post-hoc pairwise test results of the model fit differences between Stimuli at Condition level for Experiment 1 (with Tukey's adjustment): alien selection task

Contrast	Pair	Estimate	SE	z.ratio	p.value
Nonflipped	Gili variant - Nulu variant	1.35e+09	4.41e+09	6.46	<0.001
Flipped	Gili variant - Nulu variant	4.69e+08	1.42e+09	6.58	<0.001
Gili	Flipped - Nonflipped	1.00e+00	1.00e+00	-0.09	0.99
Nulu	Flipped - Nonflipped	3.00e+00	5.00e+00	0.50	0.96

Experiment 2

Table 8: Model results for suffix and alien selection in Experiment 2: Response $\sim$ Stimulus + (1+Version+Stimuli|ParticipantID) + (1|Word)

Suffix selection	Fixed effects	Estimate	Std. Error	z value	Pr(>|z|)
Established aliens	Intercept	7.48	1.87	3.98	<0.001
	Stimulus (Nulu outfit)	-13.54	3.33	-4.07	<0.001
New aliens	Intercept	1.60	0.56	2.85	<0.01
	Stimulus (Nulu outfit)	-2.96	0.75	-3.97	<0.001
Alien selection	Fixed effects	Estimate	Std. Error	z value	Pr(>|z|)
Established aliens	Intercept	1.76	0.13	13.64	<0.001
	Stimulus (Nulu variant)	-3.58	0.17	-21.59	<0.001
New aliens	Intercept	1.14	0.16	7.20	<0.001
	Stimulus (Nulu variant)	-1.98	0.19	-10.23	<0.001

Experiment 3

Table 9: New aliens: model results of suffix and alien selection in Experiment 3: Response $\sim$ Stimulus.Condition + (Version|ParticipantID) + (1|Word)

New aliens: Suffix selection	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.69	0.21	3.25	<0.01
Gilioutfit.Nonsocial	-0.23	0.29	-0.82	0.41
Gilioutift.Social	0.74	0.31	2.41	0.02
Nuluoufit.Baseline	-1.18	0.24	-4.96	<0.001
Nuluoutfit.Nonsocial	-1.61	0.29	-5.61	<0.001
Nuluoutfit.Social	-1.37	0.28	-4.81	<0.001
New aliens: Alien selection	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	0.82	0.16	5.07	<0.001
Gilivariant.Nonsocial	0.22	0.23	0.95	0.35
Gilivariant.Social	0.27	0.23	1.17	0.24
Nuluvariant.Baseline	-1.49	0.23	-6.60	<0.001
Nuluvariant.Nonsocial	-1.71	0.23	-7.55	<0.001
Nuluvariant.Social	-1.76	0.23	-7.73	<0.001

Table 10: Established aliens: model results of suffix and alien selection in Experiment 3: Response ∼ Stimulus.Condition + (Version|ParticipantID) + (1|Word)

Established aliens: Suffix selection	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.76	0.11	6.82	<0.001 ***
Giliouotfit.Nonsocial	0.58	0.17	3.47	<0.001
Gilioutift.Social	0.83	0.17	4.83	<0.001
Nuluoufit.Baseline	-1.59	0.15	-10.34	<0.001
Nuluoutfit.Nonsocial	-2.47	0.17	-14.04	<0.001
Nuluoutfit.Social	-2.42	0.17	-13.94	<0.001
Established aliens: Alien selection	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	0.91	0.14	6.55	<0.001
Gilivariant.Nonsocial	0.47	0.17	2.83	<0.01
Gilivariant.Social	0.60	0.17	3.55	<0.001
Nuluvariant.Baseline	-1.92	0.16	-12.03	<0.001
Nuluvariant.Nonsocial	-2.42	0.17	-14.45	<0.001
Nuluvariant.Social	-2.39	0.17	-14.32	<0.001

Table 11: Model results for Experiment 3 using Stimuli and Condition as interaction terms in suffix and alien selection: all the categorical predictors are sum-coded. Model configuration: Response ∼ Stimulus*Condition + (Version|ParticipantID) + (1|Word)

Task	Fixed Effects	Estimate	SE	z value	Pr(>|z|)
Suffix selection: new aliens	Intercept	0.09	0.10	0.89	0.37
	Stimuli (Gili outfit)	0.77	0.07	10.85	<0.001
	Condition (Baseline)	0.02	0.13	0.13	0.90
	Stimuli (Gili outfit) : Condition (Baseline)	-0.18	0.10	-1.90	0.06
	Stimuli (Gili outfit) : Condition (Nonsocial)	-0.09	0.10	-0.94	0.35
Suffix selection: old aliens	Intercept	-0.08	0.05	-1.58	0.11
	Stimuli (Gili outfit)	1.32	0.05	26.16	<0.001
	Condition (Baseline)	0.05	0.07	0.69	0.49
	Condition (Nonsocial)	-0.10	0.08	-1.35	0.18
	Stimuli (Gili outfit): Condition (Baseline)	-0.51	0.07	-7.75	<0.001
	Stimuli (Gili outfit): Condition (Nonsocial)	0.21	0.07	2.90	<0.01
Alien selection: new aliens	Intercept	0.08	0.07	1.18	0.24
	Stimuli (Gili suffix)	0.91	0.07	13.82	<0.001
	Condition (Baseline)	-0.00	0.09	-0.01	0.99
	Condition (Nonsocial)	-0.00	0.09	-0.00	1
	Stimuli (Gili suffix): Condition (Baseline)	-0.16	0.09	-1.75	0.08
	Stimuli (Gili suffix): Condition (Nonsocial)	0.05	0.09	0.59	0.55
Alien selection: old aliens	Intercept	-0.03	0.10	-0.29	0.77
	Stimuli (Gili suffix)	1.30	0.05	26.29	<0.001
	Condition (Baseline)	-0.02	0.07	-0.25	0.80
	Condition (Nonsocial)	-0.03	0.07	-0.48	0.63
	Stimuli (Gili suffix): Condition (Baseline)	-0.34	0.07	-5.05	<0.001
	Stimuli (Gili suffix): Condition (Nonsocial)	0.15	0.07	2.08	<0.04

Table 12: Established aliens: model results of likert scale in Experiment 3: Rating $\sim$ Match.Condition + (1|ParticipantID) + (1|Word)

Established aliens: outfit-variant match	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.09	0.04	2.10	0.04
Match.Nonsocial	-0.02	0.06	-0.37	0.71
Match.Social	0.11	0.06	1.78	0.08
Mismatch.Baseline	-0.08	0.06	-1.46	0.15
Mismatch.Nonsocial	-0.24	0.06	-4.14	<0.001
Mismatch.Social	-0.27	0.06	-4.52	<0.001
Established aliens: speaker-variant match	**Estimate**	**Std. Error**	**z value**	**Pr()**
Intercept	0.39	0.04	9.57	<0.001
Match.Nonsocial	-0.03	0.06	-0.44	0.66
Match.Social	0.01	0.06	0.23	0.82
Mismatch.Baseline	-0.68	0.05	-13.11	<0.001
Mismatch.Nonsocial	-0.86	0.06	-15.09	<0.001
Mismatch.Social	-0.80	0.06	-14.05	<0.001

Table 13: New aliens: model results of likert scale in Experiment 3: Rating $\sim$ Match.Condition + (1|ParticipantID) + (1|Word)

New aliens: outfit-variant match	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.16	0.08	2.10	0.04
Match.Nonsocial	0.24	0.11	2.19	0.03
Match.Social	0.07	0.11	0.65	0.52
Mismatch.Baseline	-0.38	0.07	-5.35	<0.001
Mismatch.Nonsocial	-0.45	0.11	-4.16	<0.001
Mismatch.Social	-0.46	0.11	-4.29	<0.001