Linguistic convergence to observed vs. expected behavior in an alien-language map task

Lacey Wade[1] and Gareth Roberts*[1]

[1]Department of Linguistics, University of Pennsylvania
Affiliation

Abstract

Individuals shift their language to converge with interlocutors. Recent work has suggested that convergence can target not only observed, but also *expected*, linguistic behavior, cued by social information. However, it remains uncertain how expectations and observed behavior interact, particularly when they contradict each other. We investigated this using a cooperative map task experiment, in which pairs of participants communicated online by typing messages to each other in a miniature "alien" language that exhibited variation between alien species. The overall task comprised three phases, in each of which participants were told that they would be paired with a different partner. One member of the pair was given explicit linguistic expectations in each phase, while the software controlled whether or not observed behavior from their partner would be consistent or inconsistent with these expectations. The other participant was given no such expectations, allowing us to control for the role of expectation. Participants converged to both observed and expected linguistic behavior, and convergence was boosted when observation and expectation were aligned. When expected and observed behavior were misaligned, participants updated their expectations, though convergence levels did not drop. Furthermore, participants generalized what they learned about one partner to apparent novel partners of the same alien species. We also discuss individual variation in convergence patterns and the lack of a relationship linguistic convergence and success at the map task. Findings are consistent with observations outside the laboratory that language users converge toward expected linguistic behavior. They also have broader implications for understanding linguistic accommodation and the influence of social information on linguistic processing and production.

Keywords: linguistic convergence; accommodation; sociolinguistics; map task; artificial language; experimental semiotics; interaction; dialogue

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

Introduction

People often shift their language in response to their interlocutors. This process, known as *accommodation*, encompasses convergence (when language shifts to become more similar to an interlocutor) as well as divergence (where it becomes less similar) and maintenance (where no shift occurs). The phenomenon of convergence is often commented upon by language users themselves and has been the focus of a great deal of research (e.g., Babel, 2009, 2012; Garrod & Doherty, 1994; Garrod & Pickering, 2004; Giles, Coupland, & Coupland, 1991; Goldinger, 1998; Nielsen, 2011; Pardo, 2006). There have been two main lines of inquiry with regard to linguistic convergence, the first focusing on convergence as an automatic, mechanistic process (e.g., Goldinger, 1998; Goldinger & Azuma, 2004; Pickering & Garrod, 2004; Trudgill, 2008) and the second focusing on the socio-psychological motivations for convergence (e.g., Bell, 1984; Bourhis & Giles, 1977; Giles et al., 1991; Shepard, Giles, & LePoire, 2001). While the former has tended to concentrate primarily on observed linguistic behavior as the cue to convergence (e.g., Garrod & Doherty, 1994; Shockley, Sabadini, & Fowler, 2004), the latter line of research has considered a wider range of cues, including non-linguistic information such as social stereotypes and personas, which may trigger linguistic expectations that language users converge toward (e.g., Auer & Hinskens, 2005; Bell, 1984, 2001; Wade, 2017, 2020). The purpose of this paper is to present an experimental study in which we manipulated the presence or absence of expected and observed linguistic behavior, along with whether or not they aligned, and then measured convergence to each.

Automatic models of convergence tend to assume tight perception-production relationships, such that the target a person converges towards (e.g., a phonetic feature) is precisely what they perceived (the *cue* to convergence). These models therefore consider observed behavior to be both the cue to and target of convergence. One such proposal for automatic mechanisms underlying convergence is the Interactive Alignment Model (Pickering & Garrod, 2004). Under this model, linguistic convergence is the result of automatic structural priming mechanisms that align on all levels of linguistic structure. Another proposal comes from work by Goldinger (1998), in which participants completed a shadowing task, reading English words aloud then repeating the same words after a model talker. Shadowed tokens were judged by naive listeners to be more similar to the model talker's productions than baseline tokens. Goldinger posited that the mechanism underlying this phonetic convergence was related to episodic memory storage and that there is a strong link between perception of linguistic forms and production of those same forms. Under an episodic account, a memory trace of the prime word produced by the speaker is activated when the listener goes to produce that word, automatically informing their subsequent production.

There have been proposals under such frameworks that treat convergence as an automatic consequence of perception, suggesting that convergence can target *only* observed behavior. For instance, Goldinger and Azuma (2004) argued that phonetic convergence does not generalize beyond observed linguistic forms; this was based on evidence that speakers imitated various phonetic properties of lexical items they had heard, but did not extend this to unheard lexical items containing the same phonemes. Other studies have found evidence, however, that imitation can generalize beyond the word level. Pardo (2006) for instance, found that speakers' productions became more similar to their conversational partners' on words that they had not heard in the exposure phase, suggesting that convergence can be

generalized across words at the phoneme level. Others have provided evidence that phonetic convergence can even generalize to new phonemes. For instance, Nielsen (2011) found that, when participants were exposed to artificially lengthened voice-onset time (VOT) for /p/-initial words, not only did participants imitate artificially lengthened VOT for new /p/-initial words, but they also produced lengthened VOT for new /k/-initial words, though the effect was somewhat weaker. She suggested that the target of imitation in this case may be at the level of a shared feature below the level of the phoneme.

Zellou, Dahan, and Embick (2017) further suggested that listeners are influenced by a mental model of the talker in shadowing tasks, as opposed to individual instances of the linguistic form itself. They found that participants who heard a hyper-nasalized speaker in the first block of a shadowing task increased their degree of coarticulatory nasalization. However, if hyper-nasalization was heard in a second block (that is, after an initial block of regular nasalization) participants' degrees of nasalization leveled out as if they were averaging the nasalization across all tokens they had heard from the speaker and converging toward that average. The authors suggested that participants may imitate isolated phonetic forms immediately after hearing them but that, after more exposure (or after a delay), they converge toward a model of the speaker based on accumulated utterances, rather than toward the most recent tokens.

Although several of these studies have proposed that the *target* of convergence may not be identical to the convergence *cue*, they all assume that the target is at least a generalization or abstraction of some cue actually present in observed linguistic behavior. Conversely, socio-psychological accounts of convergence, many of which treat convergence as a form of style-shifting occurring in response to individuals present, allow for cues beyond linguistic observation. In these accounts, language users might converge toward targets that are based on their *expectations* about their interlocutors, rather than what their interlocutors actually do.[1] Communication Accommodation Theory (CAT) suggests that the motivation for shifts in communication is to win approval or maintain social distance (Giles et al., 1991; Giles, Robinson, & Smith, 1979). CAT also accounts for non-convergence behaviors such as linguistic divergence and maintenance, as well as for hyper-convergence, where a language user overshoots the convergence target (Giles, 1980). As Bell (1984) noted, accommodative style shifting is necessarily social, as "Variation on the style dimension within the speech of a single speaker derives from and echoes the variation which exists between speakers on the 'social' dimension" (p. 151). Bell extended CAT to allow for language use to be influenced by not only one's primary addressee, but also by others who may be involved in the communicative situation (auditors, overhearers, eavesdroppers), as well as other non-personal variables such as topic and setting. His model of audience design assumes that, when speakers design their speech (including their speech style), they take into account primarily those who would be listening to that speech.

Socio-psychological accounts in which convergence results from social motivations rather

[1] Although this literature often frames convergence as a socially-influenced process under conscious control, we take the position that convergence based on social factors need not be completely controlled. Campbell-Kibler (2016), for instance, argued that "social" need not imply consciousness providing evidence that many observed social cognition behaviors must occur very rapidly and without conscious awareness. In fact, recent studies on social cognition have found that social information can influence linguistic processing in as little as 200–300 ms (Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008).

than automatic perception-production linkages are better equipped to allow for convergence toward linguistic expectations as opposed to observed linguistic behavior. For instance, Auer and Hinskens (2005) advocated an "Identity-projection model" of convergence, which suggests that instead of converging toward "observable behaviour of the recipient ... speakers converge to a stereotype of the 'model' receiver, not the actual partner in direct communication" (p. 341). That is, language users may change the way they communicate to match how they believe their conversational partner does, regardless of whether their conversational partner actually uses such linguistic forms. As evidence, they referred to Bell (2001), who reported a case of an Anglo interviewer in New Zealand frequently using the tag *eh* when conversing with a Maori interviewee but not when conversing with an Anglo interviewee. Although the Maori speaker did not use the tag *eh*, it is a feature stereotypically associated with Maori speech. This was cited as evidence that speakers shift their speech to align with beliefs about what an interlocutor *should* sound like. Accounts of expectation-based convergence can be traced further back to Thakerar, Giles, and Cheshire (1982), who termed convergence which "responds to what the speaker mistakenly assumes will be the addressee's speech on the basis of the addressee's nonspeech attributes" to be "subjective accommodation" (Bell, 1984, p.168). Earlier models of convergence in sociolinguistics such as Communication Accommodation Theory (Giles, 1980) and Audience Design (Bell, 1984) suggest that ideas about an interlocutor's identity are equally—if not more—important in eliciting convergence than the actual linguistic forms themselves. This has been more recently observed in controlled laboratory studies, in which Wade (2017, 2020) found that individuals produced more monophthongal variants of /aɪ/ (the vowel in *prize*), a salient feature of southern U.S. English, after hearing a southern-shifted model talker who never produced any tokens of /aɪ/ throughout the course of the experiment. Others studies have found similar effects but have tended to consider results to stem from topic-based style shifting or a social concept priming linguistic features rather than interlocutor-triggered convergence (e.g., Drager, Hay, & Walker, 2010; Love & Walker, 2013; Sanchez, Hay, & Nilson, 2015).

There is strong evidence that language users converge linguistically to both observed and unobserved linguistic behavior. Several studies have found convergence in the absence of strong social cues (Babel, 2010; Goldinger, 1998; Nielsen, 2011; Shockley et al., 2004) or to linguistic variants that aren't readily attached to social characteristics (such as Zellou et al. 2017, who found convergence to hypernasalization, or Nielsen 2011 and Shockley et al. 2004, who found convergence to VOT); these provide clear evidence that convergence can target linguistic features in the absence of any social information. However, findings like those of Wade (2017, 2020) and Bell (2001) provide clear evidence that social factors can trigger linguistic convergence in the absence of *linguistic* information. Importantly, observed behavior may not necessarily align with expected behavior and may, in some cases, contradict it. In such cases, Auer and Hinskens (2005) seemed to believe that social-based convergence takes precedence. They suggested that perhaps "speakers do not actually wish to conform to their co-participants' behaviour, but rather, to some kind of stereotype that they have..." (p. 342) and remarked that certain behavior such as convergence to the Maori interviewee using eh-tags flat-out contradicts models of convergence that restrict convergence cues to observed linguistic behaviors. However, the interaction between observed linguistic convergence and socially-triggered convergence has not to our knowledge

been empirically investigated. The goal of the present study is to do just that by testing how participants converge when expected and observed interlocutor behaviors fail to align. Our research questions can be outlined as follows:

1. Do participants converge toward expected linguistic behavior, in the absence of observed behavior?
2. Do participants converge toward observed linguistic behavior, in the absence of prior sociolinguistic expectations?
3. If observed and expected behavior are aligned, does this boost convergence rates?
4. How do participants reconcile convergence to contradictory observed and expected behavior?

Based on the literature we have outlined above, we predicted that participants would converge toward both observed and expected interlocutor behavior. Further, based on findings in the social psychology literature that social stereotypes are readily updated based on new information (e.g., Weber & Crocker, 1983), we predicted that expectation-based convergence would be adjusted in response to observed behavior. If that behavior was aligned with expectation, we predicted that convergence rates would be boosted. If observation contradicted behavior, however (e.g., if a speaker expecting t-flapping encountered t-glottalization instead), we predicted that convergence to the expected form would decrease. This might occur in one of two ways: Participants might revert back to a non-convergent form that is their own default linguistic form, or they might shift to converge toward the newly observed variant, possibly to a lesser extent than to the expected one. We also predicted that participants would update their expectations, becoming less likely to expect non-observed forms and creating new expectations for the recently observed form in later interactions.

To test these predictions, we ideally require a method in which participants engage in linguistic interactions (so we can measure their linguistic behavior and how it changes over the course of their interaction), which feel like genuine interactions rather than simply responses to prompts (to ensure a reasonable degree of external and ecological validity), and in which both participants' expectations of their partners' linguistic behavior *and* the observed reality of those partners' behavior can be manipulated independently. This is not trivial to achieve. The vast majority of experimental approaches in sociolinguistics involve the direct analysis of real-world linguistic behavior—investigation, that is, of the language and linguistic attitudes that subjects bring with them to the laboratory. While this engagement with authentic natural-language data is very important (and, indeed, constitutes the bedrock of the discipline), it brings with it the disadvantage that it is hard to control for variation in the linguistic background of the participants. Furthermore, if participants are to interact directly in the laboratory, it is very hard to control and manipulate the occurrence of linguistic variants in real time. The embedded nature of linguistic features, moreover, means that manipulating any one variant independently of the network of variants it inhabits is hard or even impossible to do cleanly (Weinreich, Labov, and Herzog, 1968; see Roberts and Sneller, in press, for a discussion of this in the context of artificial-language experiments). For these reasons, a few researchers have introduced artificial elements into the lab. Docherty, Langstrof, and Foulkes (2013), for instance, trained participants on a

novel sociolinguistic distribution of a familiar variable, while Rácz, Hay, and Pierrehumbert (2017) had participants learn entirely novel languages (see also Lai, Rácz, & Roberts, 2019, in press). Others have employed interactive games to create social worlds for artificial languages to inhabit. Roberts (2010) had participants negotiate in groups of four to trade resources and measured the extent to which their artificial language diverged into dialects, while Sneller and Roberts (2018) had participants play a game in which they could choose between trading with and fighting each other, and observed rates of borrowing between pre-established dialects. In these studies, which involved text-based communication, participants simply exchanged written messages freely in the artificial language. However, it is relatively simple in such experiments to have the server edit messages in-transit, giving experimenters control over how participants' language appears to other participants, while letting them believe that they are unconstrained (cf. Galantucci & Roberts, 2014; Mills, 2011; Roberts, Langstein, & Galantucci, 2016). This is the approach we took in our study, and it has several advantages. Not only does it allow relatively easy manipulation of participants' linguistic behavior, but it also simplifies data gathering and analysis. On the other hand, it is important to note that it involves a shift to the written modality, while the vast majority of work on linguistic convergence is concerned with spoken language. The difference should not be overstated—online discourse is in a number of respects rather similar to speech (McCulloch, 2019)—but there has been relatively little work on accommodation and convergence in online written interactions. What there is suggests that convergence (whether short- or long-term) occurs in broadly similar ways in online written discourse as in spoken discourse (Cassell & Tversky, 2005; Pavalanathan & Eisenstein, 2015; Pérez-Sabater, 2017). Earlier experimental work with artificial-language interactions also found strong evidence of linguistic convergence that did not look especially dissimilar to natural-language convergence (Roberts, 2010; Sneller & Roberts, 2018), except to the extent that artificial-language paradigms may amplify rates of change (Roberts, 2017). To our knowledge, however, a direct quantitative comparison has not yet been conducted on convergence in speech and writing, or between convergence in artificial languages and natural languages.

Unlike in the other artificial-language studies cited here, we had participants engage in a cooperative map task, a well-defined experimental task that requires linguistic communication between pairs of participants, while constraining the words and constructions they need to use; it has been used previously to investigate convergence in natural language (Anderson et al., 1991; Pardo et al., 2018). While participants took part in pairs online, we manipulated their expectations by leading them to believe that they were interacting with a series of different individuals, and provided explicit linguistic expectations. We manipulated observed linguistic behavior by having the server swap specific letters for other letters to either confirm or contradict expectations. We then analyzed participants' actual behavior in the dialogues to measure the extent to which they converged toward each other and how this changed over the course of their interactions.

Method

Overview

Pairs of participants completed an online map-task, in which they took turns leading each other from a starting point to a destination on a series of maps (Fig. 3). Participants

could communicate by typing messages to each other as in instant-messaging software. However, they were not permitted to use any natural language to communicate but were instead taught a miniature "alien" language to use. Each of the two participants in a pair was assigned to a different alien species—either the Greebits or the Bulbenes—and the language they were taught varied slightly depending on species (one "dialect" having [p] where the other had [f][2]). The partner assigned to be a Greebit (henceforth the "explicit-expectation participant") was given information about dialect differences that the Bulbene participant (henceforth the "no-expectation participant") was not.

The map task consisted of three phases, each containing six maps. The phases served as three separate within-subjects experimental conditions and differed from each other with respect to the explicit-expectation participants' expectations about their partner and the linguistic behavior they observed from their partner. (In what follows we will use the terms "phase" and "condition" somewhat interchangeably.) Participants were led to believe that they had a new partner for each phase and, at the start of each phase, were told the species of that partner. In fact it was always the same partner. As stated, the explicit-expectation participant was also given information about what to expect from their partner's language in each phase (either [f] or [p], depending on the species they believed their partner to be). However, what they actually observed ([f], [p], or a completely novel variant [v]) was manipulated; in particular, the server automatically swapped certain letters in their partner's messages for another letter, according to the condition. This allowed us to vary participants' expectations and experience independently, while being able to control for individual variation and to measure participants' language use and the extent to which they converged based on expectations and experience. We therefore distinguish between three types of linguistic behavior in the experiment. **actual** behavior is what that partner in fact typed before the server did anything to it; **observed** behavior consists of what a participant would receive from the server as a message from their partner; **expected** behavior is what the explicit-expectation participant was led explicitly to anticipate from their partner based on what they were told about them. All three of these might in some cases coincide; in other cases they would be different.

Participants

120 native English speakers were recruited to participate in the study in pairs for course credit or $11. One pair was excluded for giving map-task instructions in English instead of the alien language.[3] Five pairs were excluded for completing less than one full phase of the experiment. The data from the remaining 108 participants (32 from the University of

[2]For convenience we will use square brackets to represent variants of the different consonants in the language, even though the language was written, not spoken out loud.

[3]Participants were excluded for using English only if the English was used to accomplish the goals of the map task (e.g., "Go West around the Mountain.") Some pairs used occasional English for other purposes and were not excluded. In total, 12 participants (or nine dyads) used some English. Of these groups, four used English only at the very beginning of the game to ask whether their partner was there or to check that the game was working (e.g., "test", "Hello?", "Any instructions?"). One participant used English just once to provide the feedback, "wtf", to their partner halfway through the task, and another participant used English only to indicate that their partner should "submit". Three pairs used English a little more often, but were still included because their use of English was tangential to the map task itself (e.g., "now", "ok", "can you see my map?" or "idk how to delete.").

Pennsylvania subject pool, 16 of whom participated in the lab, 16 of whom participated online; 76 from the online Prolific Academic platform) are included here. Sixty-four participants (59%) reported demographic data.[4] Of these, 38 identified as female and 26 as male, and the mean age was 23.6.

Procedure

After reading the instructions, each of the two participants logged on to the game and was assigned to a different alien species, either the Greebits or the Bulbenes (Fig. 1). They were then introduced to the alien language, which consisted of 14 words. These were displayed to them as in Fig. 2 and participants were given 2 minutes to study them. After this 2-minute period, each participant practiced the alien language by translating English map directions (e.g., "Go south around the yellow hut.") into the alien language, with the wordlist still present to refer to. There were ten such sentences to translate. After each one, they would receive feedback on their translation before the next sentence appeared. If the translation was accurate in terms of spelling and word order (punctuation and capitalization were ignored, as these were permitted to be used freely in the actual map task), they were told so; otherwise they were shown the correct translation. After all ten sentences had been translated, the wordlist disappeared and participants were given five further sentences to translate. This completed their training in the alien language.

Each species learned a slightly different "dialect" of the alien language. In particular, the Greebit participants' dialect contained many words with [p] in them (Fig. 2), which always corresponded to [f] in the Bulbene dialect. Participants were told in the instructions that their conversational partners during the task might be either Bulbenes or Greebits, and that this could vary throughout the game. Importantly, they were also led to believe that they would interact with three different partners over the course of the game. This required encouraging them to think that there were at least four participants involved. (In fact, there were always only two.) For both online participants and laboratory participants, this impression was encouraged at the start of the game by displaying a screen that stated what percentage of participants had already logged on. When neither participant had logged on, the screen stated that "50%" of participants had done so; this percentage increased by "25%" for each of the two real participants. For laboratory (as opposed to online) trials the illusion was further bolstered by telling participants that they would be interacting with other people who might be in the same lab as them or could be online. Where possible we brought in two pairs of participants at once, implying that they were participating as one group, whereas in reality the two pairs were participating in parallel.

In every trial the explicit-expectation participant was given information about dialect differences that the no-expectation participant was not. In particular, a note was included with their word list stating that "The Bulbenes may speak a slightly different dialect and may use 'F' in place of 'P'." The no-expectation participant's wordlist contained no such

[4]The demographic data were gathered in a survey at the end of the experiment. It appears that a substantial minority of participants missed the link to this survey. All of those who did begin the follow-up survey seem to have completed it, suggesting that the issue was indeed due to missing the survey link rather than dropping out due to a lack of engagement or time constraints. We have no strong reason to expect that the age and gender distribution of those who did not complete the experiment differed significantly from those who did.

note. This allowed us to control for the effects of expectation on convergence behavior.

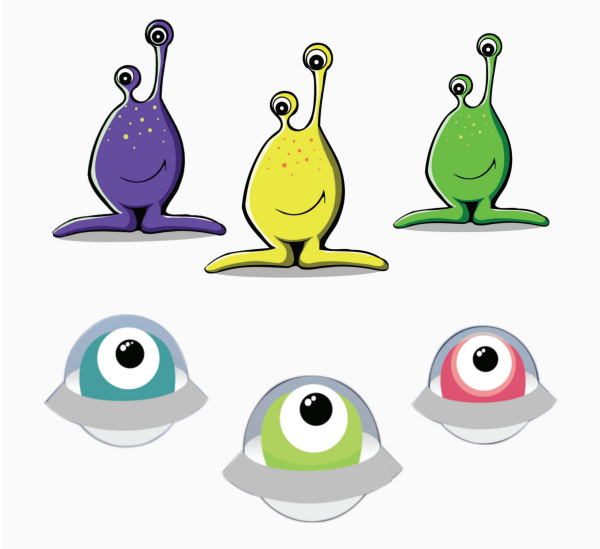

Figure 1. Greebits (top) and Bulbenes (bottom)

After the training, participants began the main map task, in which they took turns leading each other around maps. In each case, the *leader* would see a map with a goal and a route marked on it (e.g., Fig. 3). The *follower* would see the same map, but with no goal or route. The follower could use the mouse to trace a route on their map, which would be made visible to both leader and follower in real time; the leader's task was to describe the route to the follower. They could communicate by typing messages to each other using instant-messaging software. They were, however, required not to use any natural language, but to use only the alien language they had learned. The instant-messaging software was bidirectional, so either participant could send the other a message at any time during the map task. The map task was not timed, and it was not stated or implied to participants that their speed in completing each map mattered. (See Appendix for the instructions given to participants.) Each map would be considered complete only when the follower clicked a button to confirm that they had reached the goal. Then a new map would appear, and the two participants would swap roles.

There were 18 maps in total, and the game was broken into three phases, with six maps in each phase. For each phase, both participants were led to believe they would be interacting with a new partner (while in fact it was the same person). The explicit-expectation participant was always the first leader in every phase so that we could record their linguistic behavior before they had been exposed to their partner's linguistic behavior in that phase. The task took participants an average of 3 minutes per map (a total of 55 minutes for the entire map task, including partner swaps between rounds).

The phases served as different conditions that varied in terms of what species participants believed they were interacting with. This was done very simply by, at the start of the phase, announcing to each participant that they would be interacting with either a Greebit or a Bulbene, with a picture of the alien in question. For the explicit-expectation participant—who had been given information about variation in the alien language—this

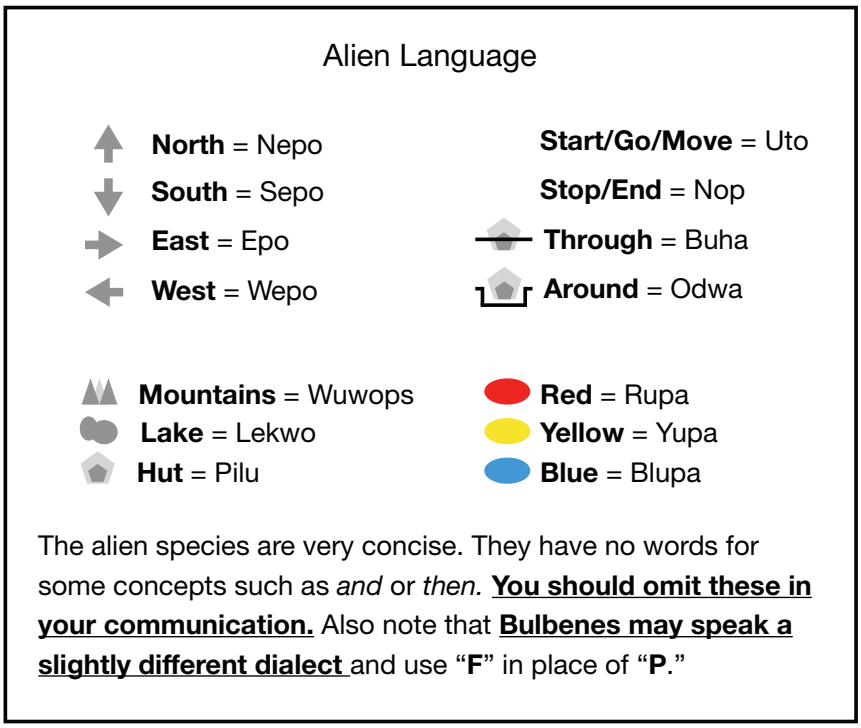

Figure 2. Example Alien Language screen, as seen by explicit-expectation participants. No-expectation participants would have seen the same language, but with [f] in place of [p], and without the final sentence about dialectal variation.

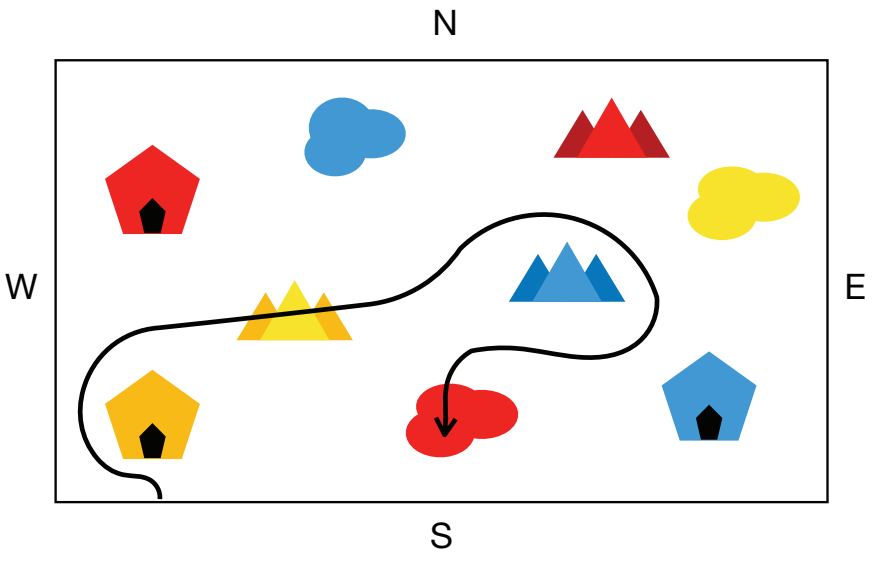

Figure 3. Example map with goal

Table 1
Summary of Experimental Design

	Explicit-expectation participant Told they are Greebit Learns language with [p]	**No-expectation participant Told they are Bulbene Learns language with [f]**
Matched	Told partner is a Bulbene who may use [f] Sees [f]	Told partner is a Greebit Sees [p]
Unmatched	Told partner is a new Bulbene who may use [f] Sees [v]	Told partner is a new Greebit Sees [p]
Same-species	Told partner is a new Greebit Sees [p]	Told partner is a new Greebit Sees [p]

entailed a specific expectation about their partner's language, which was reinforced by an explicit reminder at the start of the phase about potential dialect differences, stating "Remember, Bulbenes may speak a slightly different dialect and use F in place of P."

Even if the participant had already interacted with a species before, they were told that this would be a new member of that species. Conditions differed not only with respect to the dialect that a partner was *expected* to use; they also differed with respect to the dialect each participant would *appear* to use and thus the variants that their interlocutor would observe. This was controlled by the software, so that every [p], [f], or [v] used by a participant would be automatically changed according to condition before it reached their partner; for instance, in one condition, every [p], [f], or [v] that the no-expectation participant produced would be changed to [f] before the explicit-expectation participant player saw it (Table 1). This was done only to the message that the partner received; the participant sending the message would see it precisely as they typed it, and the messages typed were recorded in results in this form, so we could analyze participants' actual linguistic behavior.

In all three conditions, the no-expectation participant was told that they were paired with a Greebit, and the software always ensured that any [f] in their partner's messages would be turned into a [p], consistent with the Greebit dialect. For the explicit-expectation participant, however, each condition differed from the other with respect to expectation or actual experience. In the **Same-species condition**,[5] the explicit-expectation participant was told that their partner was also a Greebit. (The no-expectation participant was still led to believe that they themselves were a Bulbene and were never made aware that their partner had been given information that conflicted with this.) The software ensured that every message the explicit-expectation participant saw contained [p] in place of [f] (or [v]; see below), consistent with expectation. In the **Matched Other-species condition**, the explicit-expectation participant was led to believe their partner was a Bulbene, who might use [f] in their dialect; the software ensured that they would indeed encounter [f] instead of [p] or [v]. In the **Unmatched Other-species condition**, the explicit-expectation participant was again told they would be partnered with a Bulbene, who might use [f] in their

[5] All conditions are named based on the explicit-expectation participant's experience; as far as participants were concerned, phases had no names.

dialect. This time, however, the software confounded this expectation by changing every [f] or [p] in their partner's messages to [v]. The order of conditions was counterbalanced (Table 2), to allow us to control for and compare the influence of phase order. Note also that half of the participants were in the Matched Condition first, while the other half were in the Unmatched Condition first. This design allows us to investigate whether explicit-expectation participants generalized what they learned from one Bulbene partner's observed linguistic behavior to a novel Bulbene partner in a later phase.

After all three phases had been completed, participants were directed to a Qualtrics survey, which collected demographic information and asked participants what they believed the experiment to be about.

Table 2
Condition order counterbalanced across participant pairs

	A	**B**	**C**
Phase 1	Matched	Unmatched	Same-species
Phase 2	Unmatched	Same-species	Matched
Phase 3	Same-species	Matched	Unmatched

	D	**E**	**F**
Phase 1	Matched	Unmatched	Same-species
Phase 2	Same-species	Matched	Unmatched
Phase 3	Unmatched	Same-species	Matched

Analysis

All messages sent were recorded, allowing us to measure the degree to which participants converged linguistically with each other. Paths traced on maps were recorded (at a sampling rate of 6.7 frames per second), allowing us to measure the distance between participants' paths and the target routes, as a metric of communicative success.

Data for pairs who failed to complete a phase were omitted for that phase. For instance, if a pair completed all of phases 1 and 2, but only part of phase 3, only the data for phases 1 and 2 are included. In total, 4 pairs failed to complete one of the three phases (Order B, C, D, and F), and two pairs failed to complete two phases (Orders B and F). Each word was coded for whether it contained an [f], [p], or [v], and only words containing one of those three letters were included in the analysis. Mixed effects logistic regression models were fit using the lmerTest package in R (Kuznetsova, Brockhoff, & Christensen, 2017). The first model (Table 3) includes data only from the explicit-expectation participants and predicts usage of f (1) or another variant (0). The model includes by-speaker and by-item (each map) random intercepts. Because the explicit-prediction participant was always the leader and sent the first message, we were able to distinguish forms produced before exposure to their partner's language from forms produced after it.

Fixed predictors are as follows:

- Condition: Categorical predictor with levels Matched, Unmatched, and Same-species; treatment coded with Same-species as the reference level.

- PrePost: Categorical predictor with levels Before and After (i.e., whether word was used before or after observing partner's language); treatment coded with Before as the reference level.
- Phase: Continuous predictor referring to the phase number (1, 2, 3)
- WhichFirst: Categorical predictor with levels Unmatched and Matched, referring to which of the two conditions was first; treatment coded with Matched as the reference level.
- Two-way interaction between Condition and PrePost
- Two-way interaction between Phase and WhichFirst

Table 3

Logistic regression model predicting usage of f (1) or another variant (0) for explicit-expectation participants

	AIC	**BIC**	**logLik**	**deviance**	**df.resid**	
	2148.6	2223.1	-1063.3	2126.6	6464	
Scaled residuals	**Min**	**1Q**	**Median**	**3Q**	**Max**	
	-8.134	-0.097	-0.009	-0.001	109.064	
Random Effects	**Groups**	**Name**	**Variance**	**Std.Dev**		
	Participant	(Intercept)	60.711	7.792		
	Map	(Intercept)	.712	0.844		
Fixed effects	**Estimate**	**Std. Error**	**z value**	**p-val**		
(Intercept)	-17.915	2.634	-6.802	1.03e-11	***	
Condition(Matched)	5.420	0.552	9.823	<2e-16	***	
Condition(Unmatched)	4.793	0.391	12.257	<2e-16	***	
PrePost(After)	-0.347	0.364	-0.953	0.340		
Phase	2.198	0.285	7.717	1.19e-14	***	
WhichFirst(Unmatched)	4.337	2.176	1.993	0.046	*	
Condition(Matched):PrePost(After)	1.284	0.418	3.072	0.002	**	
Condition(Unmatched):PrePost(After)	-1.258	0.415	-3.035	0.002	**	
Phase:WhichFirst(Unmatched)	-2.756	0.484	-5.693	1.25e-08	***	

The second model (Table 4) includes data only from the no-expectation participants and predicts usage of accommodative variant p (1) or another variant (0). The model includes by-speaker and by-item random intercepts and includes only the continuous predictor Phase (1, 2, 3). Since what the no-expectation participant observed did not vary by Condition, Phase is the only relevant predictor recorded.

Data and analysis scripts are available at https://osf.io/vbqdy/

Results

Participants seem to have learned the language well; in the final five sentences, which were translated without a key, participants got less than 2% of letters wrong. The data set for the experiment itself consists of 19,600 total recorded words, 11,825 of which contain the variable of interest, with a mean of 219 observations (sd = 89.4) per pair of participants. The [f] variant was produced 53% of the time, the [p] variant 46% of the time, and the [v] variant 1% of the time.

Table 4
Logistic regression model predicting usage of p (1) or another variant (0) for no-expectation participants

	AIC	**BIC**	**logLik**	**deviance**	**df.resid**
	2200.2	2227.1	-1096.1	2192.2	6147
Scaled residuals	**Min**	**1Q**	**Median**	**3Q**	**Max**
	-5.148	-0.187	-0.091	-0.039	13.706
Random Effects	**Groups**	**Name**	**Variance**	**Std.Dev**	
	Participant	(Intercept)	7.103	2.665	
	Map	(Intercept)	0.237	0.487	
Fixed effects	**Estimate**	**Std. Error**	**z value**	**p-val**	
(Intercept)	-5.774	0.519	-11.115	<2e-16	***
Phase	0.565	0.087	6.516	7.24e-11	***

Convergence to expected behavior

Our first question was whether participants converged to expected linguistic behavior, and whether this varied across conditions. We focus here only on explicit-expectation participants, as these were the only participants who began the task with sociolinguistic expectations. For our purposes, convergence to expected behavior meant that an explicit-expectation participant used the variant [f] when conversing with a partner they believed to be Bulbene, before observing their partner's linguistic usage. While they would also expect [p] from a Greebit partner in the Same-species condition, we treat this as a control condition since [p] is also the variant this group learned.

As Fig. 4 shows, explicit-expectation participants used more of the variant [f], which they expected from their Bulbene partners, before observing their partners' speech in both the Matched and Unmatched conditions. In the Same-Species condition, in which [p] would be expected, participants' pre-observation [f]-rates were only 4.4% (sd = 20.4), compared with 15.3% (sd = 36) in the Matched Condition and 29.3% (sd = 45.5) in the Unmatched Condition.[6] Before observing their partner's linguistic behavior, explicit-expectation participants used more [f] when they believed that was a feature of their partner's dialect than when they did not. The model suggests that this difference in [f] usage based on expectation alone (i.e., before these participants observed their partners' actual linguistic behavior) is significant. Because the PrePost predictor is treatment coded and put into the model as an interaction with Condition, the main effect of Condition considers only the pre-observation data. The significant main effect of Condition, then, suggests that explicit-expectation participants began the Other-Species conditions by producing more [f] than in the Same-Species condition, even before seeing any instances of the variable of interest from their partner. This effect is unlikely to have been due to learning errors, or explicit-expectation participants forgetting the form they had learned and copying their partners, as there was almost no [f] usage in the Same-Species Condition, regardless of when it occurred; furthermore, participants used [f] before observing their partner's speech (Fig. 7).

[6]The unexpected difference in pre-observation [f] usage between the Matched and Unmatched conditions is explored in Section 4.3

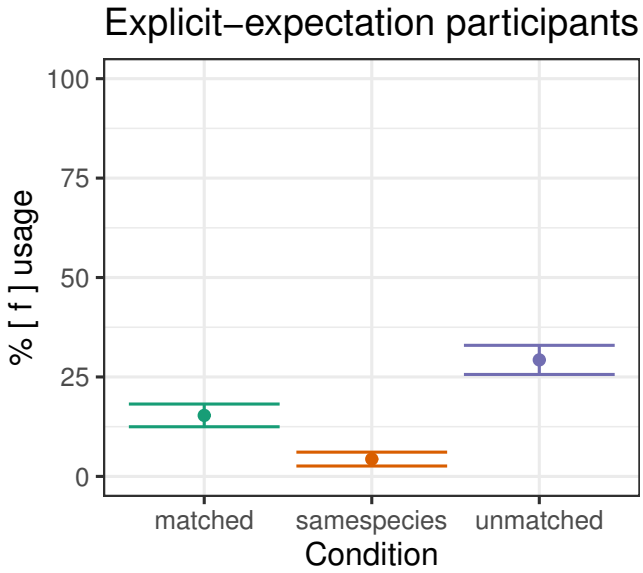

Figure 4. Explicit-expectation participants' convergence toward expected behavior. Figure shows [f] rates across conditions, with mean values for [f] usage prior to observing partners' linguistic behavior, with 95% confidence intervals.

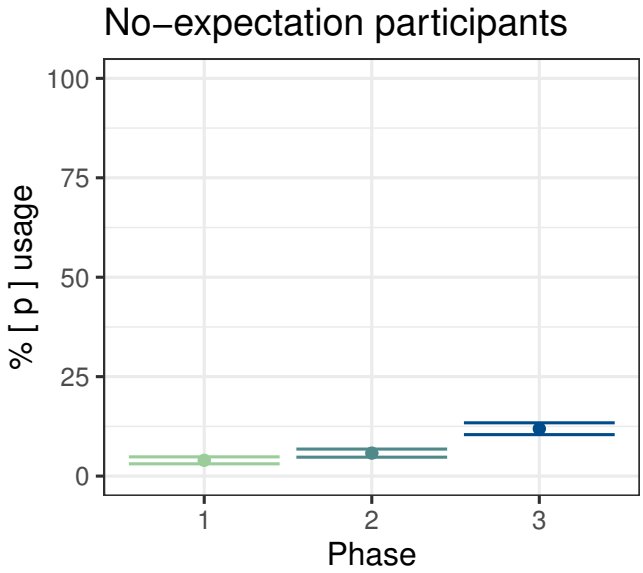

Figure 5. No-expectation participants' rates of convergence to observed [p] by phase order. Mean values with 95% confidence intervals.

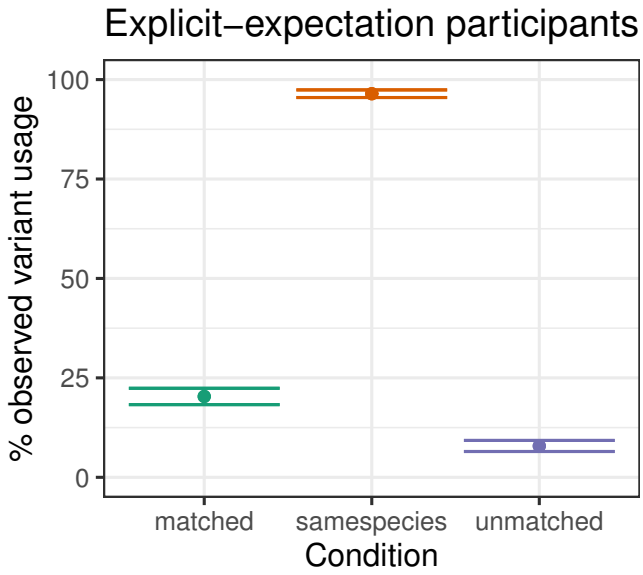

Figure 6. Explicit-expectation participants' convergence toward observed variants [f] (Matched), [p] (Same Species), and [v] (Unmatched) rates by condition. Mean values for post-observation variants with 95% confidence intervals

Convergence to observed behavior by no-expectation participants

Our second question was whether participants converged to observed linguistic behavior in the absence of prior expectations. Here the relevant data come from the no-expectation participants. As can be seen in Fig. 5, in which the phases are arranged in order of occurrence for no-expectation participants, convergence was low in the first phase but increased over the course of the experiment from a mean of 4% to a mean of 12%. The model predicting no-expectation participants' [p] usage shows a main effect of Phase, meaning that they converged significantly more to their partners' usage of [p] as the experiment progressed. This increase suggests that no-expectation participants were forming expectations about Greebits' language from their conversation partners and generalizing their experiences to (apparent) new partners. The no-expectation participants' levels of convergence are comparable to those of the explicit-expectation participants when expectation contradicted observation (explicit-expectation participants' post-observation convergence toward [v] in the Unmatched Condition was 7.9%), though explicit-expectation participants appeared to converge to a greater extent overall. This difference should be interpreted with caution, however, since the two groups were comprised of different pools of participants and—as discussed below—there was considerable individual variation between participants in level of convergence.

Convergence to observed behavior by explicit-expectation participants

Our third and fourth questions concerned convergence to observed behavior in the context of explicit expectations. As Figure 6 shows, explicit-expectation participants responded in all conditions to observed behavior, whether this contradicted or was consistent with their

expectations, including using [v] at a rate of 7.9% after observing their partner using this form, contrary to expectations. Further, observed behavior interacts with expectations. As Figure 7 shows, in the Matched condition, convergence to [f] increased by 5 percentage points to 20.3% (at the expense of [p]) after expectations had been confirmed by partner behavior. In this condition [v] use remained at nearly 0%. In the Unmatched condition, mean [f] use decreased by 8.5 percentage points after observing partner use of [v] in place of [f]. Explicit-expectation participants' use of [v] in this condition rose accordingly from 0 to 7.9%, while [p] usage remained roughly constant, suggesting that convergence rates per se remained constant, and what changed was the *target* of convergence. Convergence to [f] also did not drop to zero after experience of contradictory behavior in the Unmatched condition. As expected, there was no significant main effect of PrePost with Same-species as the reference level (Est. = -.347, $p = .34$), suggesting no shift in [f] usage from pre-observation to post-observation, when [f] would have been unexpected. However, significant interactions between Condition and PrePost suggest that explicit-expectation participants did exhibit significant shifts pre- to post-observation in the other experimental conditions. While levels of convergence remained very constant throughout the Same-species condition, participants produced significantly more [f] post-observation in the Matched Condition (Est. = 1.284, $p = .002$), and significantly less [f] post-observation in the Unmatched Condition (Est. = -1.259, $p = .002$).

A surprising result is that explicit-expectation participants' pre-observation [f] usage was higher in the Unmatched condition than in the Matched condition (29% vs. 15%), where we should not expect a difference. One possible explanation might be that there was an asymmetry between the two conditions in terms of order effects. Whenever it did not come first, the Unmatched phase would always follow a phase in which expectation matched observation. The Matched phase, by contrast, came after the Unmatched phase—in which observation contradicted expectation—in half of all trials, and directly after it in a third of trials. However, as can be seen in Fig. 8, this cannot completely explain the effect—[f] usage was lower in the Matched condition than in the Unmatched condition even when it came first. A more important explanation concerns where the experiment was conducted. Thirteen percent of participants assigned to be in the Matched condition first (Orders A, C, and D) completed the task in the lab, while only 2% of participants assigned to the Unmatched condition first were laboratory participants. This was due partly to chance, and partly to an error that caused the first four pairs of participants run in the lab to be assigned to order A, making up nearly half of the participants in this group. Participants who completed the task in the lab were apparently less inclined to converge, particularly before observing their partner's speech. Indeed, when participants who completed the task in the lab are excluded, the 14-percentage-point gap in pre-observation [f] usage between the initial Matched and initial Unmatched phases decreases 7 percentage points.

There is more than one possible reason for this effect. First, participants in the lab had an opportunity to see each other (if only briefly) before the experiment began. This may have influenced their likelihood of becoming as deeply engaged in the roleplay aspect of the game (although this is not especially consistent with other work using similar paradigms; e.g., Roberts 2010; Sneller and Roberts 2018). Second, demand characteristics may have played a greater role in the lab, such that participants with an experimenter present felt more

pressure to perform "correctly," including using the alien language exactly as presented.[7]

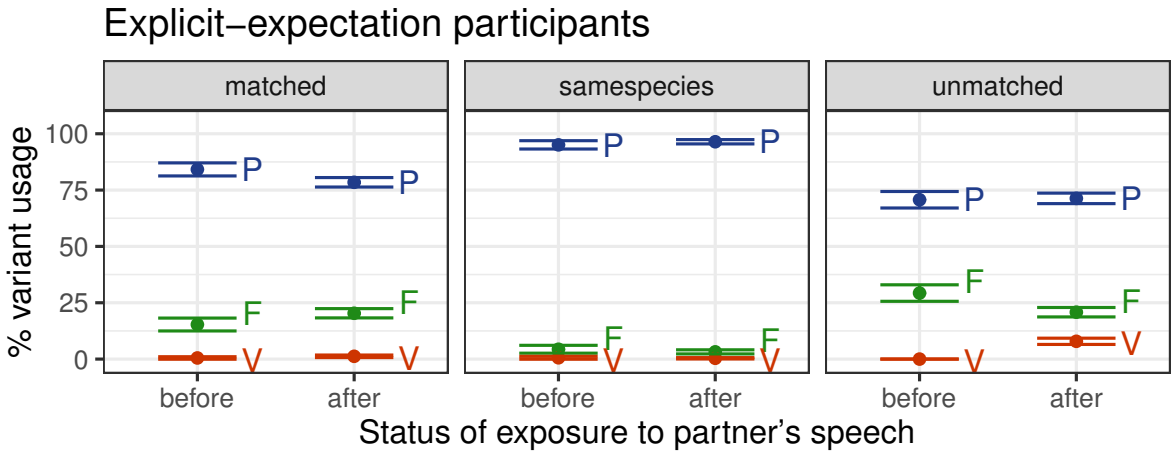

Figure 7. Explicit-expectation participants' rates of all three variants, by condition and pre-/post-exposure to partner's speech, with 95% confidence intervals.

Effects of order

For explicit-expectation participants, there was a main effect of Phase (Est. = 2.198, $p < .001$) and a significant interaction between Phase and WhichFirst (Est. = -2.756, $p < .001$). As can be seen in Fig. 8, condition order played an important role in influencing participants' subsequent behavior. Where the Matched condition came before the Unmatched condition (even if separated by the Same-species condition), the increased [f] usage in response to confirmatory partner behavior "carried over" to the pre-observation component of the Unmatched condition, (in fact, rising slightly from 21.4% to 30.4%), before falling to 20.6% in response to the new mismatch between observation and expectation. The reverse pattern occurred when the Unmatched condition came before the Matched condition, reducing pre-observation [f] usage in the latter from 21% to 16.7%, then rising to 19.3% after participants observed linguistic behavior that confirmed pre-existing expectations. However, in this case the rise in [f] usage was smaller than when the Matched condition came first, suggesting that the contradiction of expectations might have a greater and longer lasting effect than confirmation. Also note that, while convergence toward observed variant [v] increased during the Unmatched condition, this increase did not carry over to an apparently new partner in the next matched condition. Rather, [v] usage started off (and remained) relatively low in the subsequent Matched condition.

Individual differences in convergence rates

The analysis reported above masks considerable individual variation, in terms of both whether and how participants adjusted their linguistic behavior after observing their partners'. Roughly half of the explicit-expectation participants simply did not use [f] at all

[7] A similar, but distinct, explanation might be that lab participants worked harder at learning the alien language, so that online participants were more likely to forget the language and take their cue from their interlocutor. This seems unlikely, however, as it would not explain pre-observation convergence.

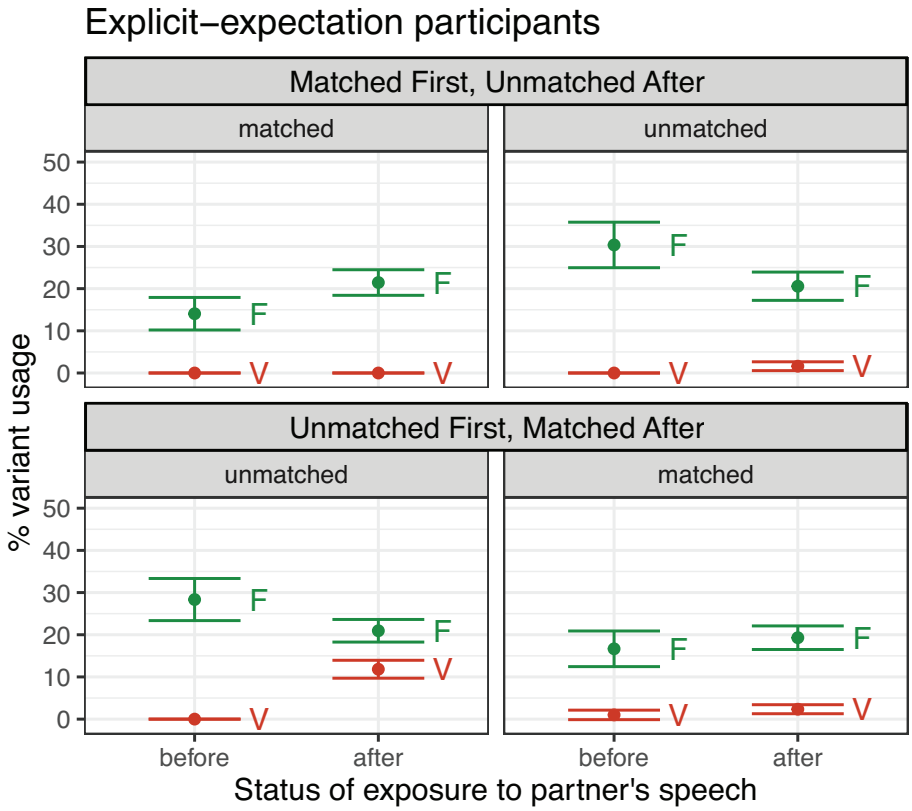

Figure 8. Explicit-expectation participants' rates of [f] and [v] pre- and post-observation of partner's speech, broken down by relative phase order. (Error bars show 95% confidence intervals.)

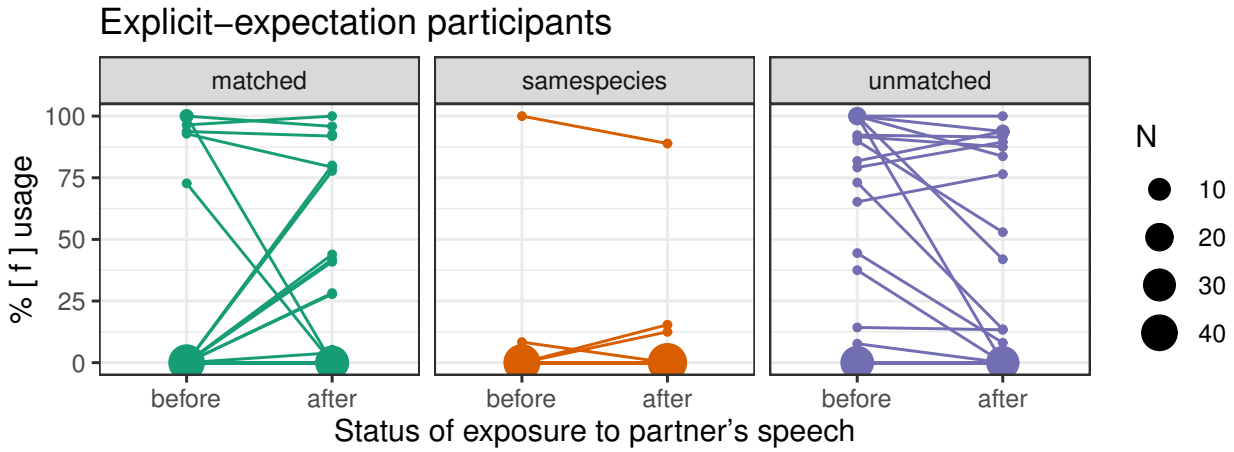

Figure 9. Individual explicit-expectation participants' patterns of adjustment after observing partners' speech, by condition. Points represent [f] rates for each explicit-expectation participant, and point size indicates the number of participants with that value.

throughout the course of the experiment (25 used only [p] and five used a mixture of [p] and [v]). Seventeen participants varied between using [p] and [f], while seven used all three variants. Nobody used only [v]. In the Same-species condition, most explicit-expectation participants used close to 0% [f] throughout the phase (N = 49). However, one participant began the phase with 100% [f], which decreased to 89% after observing their partner using [p].

In the Matched condition, the majority of participants showed no shifts (either remaining at 0% or 100% [f] throughout). 18% shifted as expected, increasing [f] usage post-observation, while 10% exhibited unexpected shifts in the opposite direction. All of the participants who exhibited large convergence shifts after observing their partner's confirmatory behavior started at or near 0% [f] (N = 5). Starting points were rather bi-modal, with participants starting at either 0% or close to 100% [f] usage. Most of those who started closer to 100% [f] usage kept a consistently high rate after exposure to partners' speech. In the Unmatched condition, most participants again exhibited no shifts, while 24% exhibited the expected downward shift in [f] usage after observing their partner using [v]. Only 6% exhibited shifts in the opposite direction. Starting and ending points varied widely in this condition (Fig. 9).

Communication style

There were considerable differences between participants in how they communicated. We distinguish three main communication styles used by participants. 27 pairs primarily employed a style in which the leader for a given map gave directions using a single chunk of text, and the follower said little. (It should be recalled that members of a pair took turns to be leader and follower, so there was still significant back-and-forth between partners within any given phase.) 15 pairs primarily employed a style in which the leader again dominated, but used multiple messages. The remaining 12 pairs employed a style in which there was continuous back-and-forth between partners over a single map, including a number of clarification questions.

We investigated whether higher levels of interaction between pairs resulted in higher rates of convergence, by recording each pair's rate of turn taking, operationalized by counting the number of adjacent pairs of messages from different participants. A pair of messages was coded as an adjacency pair if the two messages were directly adjacent and each was from a different member of the participant dyad. There was a moderate correlation between adjacency-pair rate and the no-expectation participants' usage of accommodative variant [p] (Pearson's R = 0.188, $p = .02$). However, the mean number of adjacency pairs per dyad did not predict explicit-expectation participants' convergence toward expected or observed behavior.

Success on the map task

We measured success on the map task in terms of the distance between the route line marked on the leader's map and the line drawn by the follower, the intuition being that the closer the two lines were to each other, the more successfully the pair had communicated. We measured the distance as follows. First we measured the length of each line. Then we divided each line into 10 approximately equal segments and, for each of these ten line segments,

we identified the centroid and measured the distance between that and the corresponding centroid on the other line. Our final distance metric was calculated as the mean centroid distance for the pair of lines. This was then normalized by dividing the difference between the distance and the minimum distance across dyads by the difference between the maximum and minimum distances, producing a score between 0 and 1. We then subtracted this score from 1 so that a higher value would correspond to greater success. On the whole, participants communicated reasonably well. The mean success score was 0.74 (sd = 0.14), and most dyads (60%) had scores above 0.75. The median score was 0.78. An obvious question is whether success correlated with measures of convergence or communication style. Interestingly, it did not correlate with either mean convergence for either alien species (operationalized as the mean rate of use of [f] and [v] by explicit-expectation participants and [p] by no-expectation participants) or to turn taking.

Discussion

We taught pairs of participants a miniature artificial language and had them engage in a cooperative map task in which we manipulated both what they expected of each other's linguistic behavior and what they observed. In doing so, we sought to answer four questions. First, would participants converge toward the linguistic behavior we led them to expect from their interlocutors, in the absence of observed behavior? Second, would they converge toward observed linguistic behavior, in the absence of explicit expectations? Third, would alignment between observed and expected behavior boost convergence? Fourth, would misalignment between observed and expected behavior reduce convergence? In all four cases, we predicted that the answer would be yes.

We found support for our predictions that participants would exhibit convergence to both observed and expected linguistic behavior. This is consistent with socio-psychological accounts of convergence, in which the target of convergence is not towards observable behavior so much as a mental model of individuals present, which is influenced by observed linguistic behavior, but also by stereotypes, nonlinguistic social observations, and other expectations. Our results also mirror recent laboratory studies of natural language that have observed convergence to non-observed linguistic forms in natural language (Wade, 2017, 2020). Our data also supported our broad prediction that expectation-driven convergence, as exhibited at the start of an interaction, would be adjusted to account for newly observed behavior. Observed behavior that was consistent with expectation, as in the Matched Other-species condition, led to an increase in convergence, consistent with our prediction. Furthermore, explicit-expectation participants in this condition exhibited higher levels of convergence than no-expectation participants in any phase of the experiment.[8] In the Unmatched condition, where expectation and observation were misaligned, participants increased convergence to observed behavior, at the expense of the expected variant, but overall levels of convergence in fact remained very similar. Rather than causing a reduction in convergence, in other words, misalignment merely caused a restructuring

[8]Though the latter exhibited increasing convergence over time, so it is possible that over longer time periods they would have reached or exceeded the explicit-expectation participants' rate of convergence. We should also remain aware, given the substantial individual differences observed, that comparisons of explicit-expectation participants with no-expectation participants are between-subjects comparisons, while other comparisons are within-subjects.

of targets—although notably the *expected* target did not disappear. In fact, convergence to the expected variant always remained higher than convergence to the observed variant, suggesting that sociolinguistic expectations are quite strong.

Based on social-psychological work on stereotype change (Rothbart & Park, 2003), we can hypothesize that explicit-expectation participants' reconciliation of contradictory observed and expected information might have been accomplished in at least two distinct ways. First, explicit-expectation participants might update their expectations about all Bulbenes based on observations of a single Bulbene partner's behavior; they would then apply this updated expectation of the species to new Bulbene partners. Alternatively, explicit-expectation participants might classify their expectation-contradicting partner as a non-typical member of the species, thereby "subtyping" this member, leaving expectations of the Bulbene species as a whole unchanged; we would then expect them to exhibit no change in convergence behavior with novel Bulbene partners. While individuals may have employed different strategies, the aggregate order effects we observed for explicit-expectation participants support the first account: When the Matched condition preceded the Unmatched condition, initial convergence started high in the latter; conversely, when the Unmatched condition came first, convergence in the Matched condition began at a lower rate than at the end of the Unmatched condition. Explicit-expectation participants therefore seemed to have generalized what they learned to novel talkers of the same species, suggesting a general update of their linguistic expectations for the Bulbene species. Figure 9 suggests that both strategies may have been used by participants, as some ceased [f] usage completely after observing contradictory behavior, some decreased [f] usage somewhat, while others exhibited no significant change. No-expectation participants—who were given no explicit expectations and observed the same behavior in every phase—also appeared to generalize what they learned in a given phase to subsequent apparently novel partners, exemplified by their increasing convergence to the Greebit variant [p] over the course of the experiment. This observation suggests that expectations were quickly formed within a single phase of the experiment, even in the absence of explicit awareness that species might vary in their linguistic usage.

The order effects we observed for explicit-expectation participants led to an asymmetry between conditions, with the consequence that, overall, convergence at the very beginning of the Unmatched condition was likely to be higher than at the beginning of the Matched condition. That is, the Unmatched condition was never preceded by a condition in which expected convergence was contradicted, and in most cases followed conditions in which it was consistent with observation. By contrast, the Matched condition never followed a condition in which expectations about Bulbenes were consistent with observation and in three out of six cases followed conditions in which observations were *inconsistent* with expectations. We should expect this to have led to higher rates of convergence at the beginning of the Unmatched condition than at the beginning of the Matched condition. However, while this is indeed what the results show, the data suggest that the asymmetry described cannot be the sole explanation—convergence rates were higher in the Unmatched condition than the Matched condition even when it was the first phase. As suggested above, this seems partly a result of participants who took part in the lab (and who ended up being somewhat over-represented in trials where the Matched condition was the first phase) being slower to converge than online participants, perhaps as a result of demand characteristics. On

the one hand, this is an important methodological observation, as it suggests we should be serious about taking setting into account in future work. On the other hand, we can at least feel reassured that this seems to have constituted a nuisance variable rather than a confound, as the *pattern* of results in the lab did not differ from the pattern or results for online participants. If we are right about demand characteristics playing a greater role in the lab setting, it is also encouraging that this led to convergence rates being *reduced*. This suggests that participants had not identified that we were interested particularly in convergence (or we might have expected rates to go up) and that the convergence behavior observed among online participants was relatively "natural".

A related point concerns individual differences across settings. Participants varied considerably in whether, and to what degree, they converged. This is consistent both with anecdotal observation and with observations from the scientific literature (Babel, 2012; Sonderegger, Bane, & Graff, 2017; Yu, Abrego-Collier, & Sonderegger, 2013; Zellou, 2017). It is thus reassuring to see it manifest itself in an artificial-language experiment like this, even if it introduces a further nuisance variable. It would be interesting in future work to explicitly compare intra-individual rates of convergence in natural and artificial languages.

All the same, it is interesting that levels of convergence did not seem to be related to communicative success or turn taking. One might have expected greater linguistic convergence to correlate with greater alignment in other domains, including on the map. This would be consistent with the interactive alignment model where multilevel alignment is a part of establishing common ground, which plays a role in successful communication (Pickering and Garrod 2004; this does not necessarily imply that greater phonetic or—as here—spelling alignment aids better communication; it may simply be a by-product of a broad pattern of alignment). It would also be consistent with socio-psychological accounts in which greater convergence would pattern with more positive social attitudes, which might be expected to support better communication (Bourhis & Giles, 1977; Giles, 1980). There are several possible reasons why we failed to find an effect. First, even if convergence ultimately serves communicative and social goals in general, this does not mean necessarily that the particular convergence rates we measured here need be directly related to the particular metric of communicative success we employed; it is not obvious, after all, that being more like your conversation partner with respect to the choice between [p], [f] or [v] will make all that much difference in a particular attempt to communicate a trajectory on a map. Second, it is crucial to note that convergence was to *observed*, not actual, behavior in any case. Our manipulation, in other words, prevented real mutual convergence. Attempts at alignment would generally not be noticed by a partner because the software manipulated individual's actual variant usage depending on the condition and therefore would have little communicative impact. That being said, we might still have expected that individuals who are more *inclined* to converge linguistically (as evidenced by greater convergence to observed behavior) might be better at coordination tasks. It is indeed possible that such an effect exists in our data, but is not identifiable using our line-comparison measure. However, it is also possible that there is no such relation. In this context we might consider earlier work suggesting that people do not in fact communicate more successfully with friends than with strangers (Savitsky, Keysar, Epley, Carter, & Swanson, 2011).

Our findings may have several implications for convergence in natural languages. First, we provide experimental evidence that mirrors the findings of Wade (2017, 2020) that

convergence based on expectation alone does in fact occur, supporting the idea that this is likely a real and replicable phenomenon. Additionally, while Auer and Hinskens (2005) suggested that expectation-driven convergence takes precedence over observed behavior, based on Bell's anecdote that an interviewer used the Maori-associated "eh" tag when conversing with a Maori speaker who never used this tag, we have not necessarily found support for the claim that expectation takes precedence over observation. Rather, we found that sociolinguistic expectations are updated relatively rapidly based on new contradictory evidence and suggest that perhaps observations that do not support expectations (i.e., because they are absent) influence speech less than observations that flat-out contradict expectation. This intuition that contradictory information carries more weight than non-confirmatory information could certainly be tested empirically using natural language. At the same time, we found that evidence contradictory to expectation does not seem to lead to total abandonment of prior expectations, as some participants continued to use [f] despite observing only [v] from their partner. This may reflect individual differences; some individuals may be more inclined to update expectations based on contradictory evidence, while others may need more time or input to do so. We expect that these individuals differences would also extend to real-world communication outside of the lab.

It is worth commenting, finally, on some advantages and limitations of our method. There are a number of advantages to investigating our question using a laboratory-language paradigm, and indeed using written language—not least the ability to have full control over participants' expectations and the social categories of the temporary world they inhabit, and to easily manipulate what language they are exposed to. It is also possible that the laboratory-language approach allowed us to amplify accommodation rates beyond what we might have expected from the more entrenched languages that the participants speak outside the laboratory (cf. Roberts, 2017). However, the simplifications that bring such advantages with them also have the consequence that the experiment involved a radically simple language, and social categories that were absolute and did not intersect or overlap with other social categories as in the real world. Future studies on how speakers reconcile contradictory linguistic expectations and observations might do well to investigate more nuanced social categories, as well as more complex languages that afford more complex patterns of accommodation. An important related point concerns modality—written communication is very different from spoken communication in a number of ways (although the instant-messaging medium, like other online discourse, is considerably more speech-like in many ways than other writing; McCulloch 2019). One consequence is that the written medium might have boosted the convergence rate, as converging with an interlocutor in writing requires only pressing one key instead of another, a much easier prospect than making a phonetic shift to a non-native phonetic form. On the one hand, this is an advantage, and a reason for adopting our method; on the other hand, it brings with it a reduction in ecological validity. Future work in this line should consider a wider range of modalities.

Acknowledgements

We thank Katherine Dix for designing the initial version of the software that this game ran on and Daksh Chhokra for devising the line-comparison analysis, Madeleine McGrath for supervising running trials in the lab, and several other members of the Cultural Evolution of Language Lab for running those trials.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, *34*(4), 351–366.

Auer, P., & Hinskens, F. (2005). The role of interpersonal accommodation in a theory of language change. In P. Auer, F. Hinskens, & P. Kerswill (Eds.), *Dialect change. The convergence and divergence of dialects in contemporary society* (pp. 35–57). Cambridge University Press.

Babel, M. E. (2009). *Phonetic and social selectivity in speech accommodation.* Berkeley, CA: University of California, Berkeley.

Babel, M. E. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, *39*(4), 437–456.

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

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204.

Bell, A. (2001). Back in style: Reworking audience design. In P. Eckert & J. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 139–169). Cambridge: Cambridge University Press.

Bourhis, R., & Giles, H. (1977). The language of intergroup distinctiveness. In *Language, ethnicity and intergroup relations* (pp. 119–135). London, UK: Academic Press.

Campbell-Kibler, K. (2016). Toward a cognitively realistic model of meaningful sociolinguistic variation. In A. Babel (Ed.), *Awareness and control in sociolinguistic research* (pp. 123–151). Cambridge: Cambridge University press.

Cassell, J., & Tversky, D. (2005). The language of online intercultural community formation. *Journal of Computer-Mediated Communication*, *10*(2), 1027.

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

Drager, K., Hay, J., & Walker, A. (2010). Pronounced rivalries: Attitudes and speech production. *Te Reo*, *53*, 27–53.

Galantucci, B., & Roberts, G. (2014). Do we notice when communication goes awry? An investigation of people's sensitivity to coherence in spontaneous conversation. *PloS one*, *9*(7), e103182.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*(3), 181–215.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11.

Giles, H. (1980). Accommodation theory: Some new directions. *York Papers in Linguistics*, *9*(105–136).

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, N. Coupland, & J. Coupland (Eds.), *Contexts of accommodation: Developments in applied sociolinguistics.* Cambridge: Cambridge.

Giles, H., Robinson, W. P., & Smith, P. M. (1979). *Language: Social-psychological perspectives. Selected papers from the first International Conference on Social Psychology and Language, held at the University of Bristol, England, July 1979.* Oxford: Pergamon Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Lai, W., Rácz, P., & Roberts, G. (2019). Unexpectedness makes a sociolinguistic variant easier to learn: An alien-language-learning experiment. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 604–610).

Montreal, QB: Cognitive Science Society.

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

Love, J., & Walker, A. (2013). Contextual activation of Australia can affect New Zealanders' vowel productions. *Journal of Phonetics*, *48*, 76–95.

McCulloch, G. (2019). *Because internet: Understanding the new rules of language*. New York: Riverhead Books.

Mills, G. (2011). The emergence of procedural conventions in dialogue. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 471–476). Austin, TX: Cognitive Science Society.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Pardo, J. S., Urmanche, A., Gash, H., Wiener, J., Mason, N., Wilman, S., ... Decker, A. (2018). The Montclair map task: Balance, efficacy and efficiency in conversational interaction. *Language and Speech*.

Pavalanathan, U., & Eisenstein, J. (2015). Audience-modulated variation in online social media. *American Speech*, *90*(2), 187–213.

Pérez-Sabater, C. (2017). Linguistic accommodation in online communication: The role of language and gender. *Revista Signos. Estudios de Lingüística*, *50*(94), 265–286.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–226.

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

Roberts, G. (2010). An experimental study of the role of social selection and frequency of interaction in linguistic diversity. *Interaction Studies*, *11*(10), 138–159.

Roberts, G. (2017). The linguist's Drosophila: Experiments in language change. *Linguistics Vanguard*, *3*, 20160086.

Roberts, G., Langstein, B., & Galantucci, B. (2016). (In)sensitivity to incoherence in human communication. *Language & Communication*, *47*, 15–22.

Roberts, G., & Sneller, B. (in press). Empirical foundations for an integrated study of language evolution. *Language Dynamics and Change*.

Rothbart, M., & Park, B. (2003). The mental representation of social categories: Category boundaries, entitativity, and stereotype change. In V. Yzerbyt, O. Corneille, & C. M. Judd (Eds.), *The psychology of group perception: Perceived variability, entitativity, and essentialism* (pp. 60–76). New York: Taylor and Francis Group.

Sanchez, K., Hay, J., & Nilson, E. (2015). Football versus football: Effect of topic on /r/ realization in American and English sports fans. *Language and Speech*, *56*(4), 443–460.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology*, *47*(1), 269–273.

Shepard, C. A., Giles, H., & LePoire, B. A. (2001). Communication accommodation theory. In W. P. Robinson & H. Giles (Eds.), *The new handbook of language and social psychology* (pp. 33–56). London: John Wiley & Sons Ltd.

Shockley, K., Sabadini, L., & Fowler, C. (2004). Imitation in shadowing words. *Attention, Perception, and Psychophysics*, *66*(3), 422–429.

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

Sonderegger, M., Bane, M., & Graff, P. (2017). The medium-term dynamics of accents on reality television. *Language*, *93*, 598–640.

Thakerar, J. N., Giles, H., & Cheshire, J. (1982). Psychological and linguistic parameters of speech accommodation theory. In C. Fraser & K. R. Scherer (Eds.), *Advances in the social psychology of language* (pp. 205–255). Cambridge: Cambridge University Press.

Trudgill, P. (2008, Mar). Colonial dialect contact in the history of european languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, *37*(02).

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

Wade, L. (2017, November). *Do speakers converge toward variants they haven't heard?* Paper presented at NWAV 46.

Wade, L. (2020, January). *Speakers converge toward variants they haven't heard: The case of southern monophthongal /ay/.* Paper presented at the LSA annual meeting.

Weber, R. C., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, *45*(5), 961–977.

Weinreich, U., Labov, W., & Herzog, M. I. (1968). Empirical foundations for a theory of language change. In W. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics: A symposium* (pp. 97–195). Austin and London: University of Texas Press.

Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PLoS ONE*, *8*(9), e74746.

Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics*, *61*, 13–29.

Zellou, G., Dahan, D., & Embick, D. (2017). Imitation of coarticulatory vowel nasality across words and time. *Language, Cognition And Neuroscience*, 1–16.

Appendix: Instructions for participants

Premise: You will be playing a game today. You are an alien species in an unfamiliar land. You will be assigned an alien partner, and you both take turns giving each other directions to various landmarks on this new land. Your partner may be the same species as you or a completely different one. In order to communicate with your partner, you must use an alien language, which is intelligible to you both.

Tasks: Your task as INSTRUCTION GIVER is to provide instructions that will allow your partner to trace a path as close to the original as possible. Only you will see the original path. You should specify starting points, ending points, and which landmarks you may have to go through or around to get there. You will be able to see your partner's movement on the screen as they follow your instructions. Your task as PATH DRAWER is to trace the path as carefully as you can based on your partner's instructions. You may ask clarifying questions to help your path match the original path as closely as possible. You will alternate your role as INSTRUCTION GIVER or PATH DRAWER every round.

Language: You will be given a list of words in the alien language. You are only to use these alien words when communicating with your partner. Any use of English will disqualify you. You will be given two minutes to memorize these words. You will then have a chance to practice using the alien language before the game begins. Do not copy down the words. We want to see how well you can us the alien language from memory alone.

Partner: There are two alien species, the Greebits and the Bulbenes. You will be told at the start of the game which species you are. You will have a new partner for each of the 3 rounds. Before each round, you will be told which species your partner is.

Click the READY TO PLAY! *button when you have finished reading these instructions and are ready to begin.*