Asymmetric accommodation during interaction leads to the regularisation of linguistic variants

Olga Fehér[a], Nikolaus Ritt[c], Kenny Smith[b]

[a]*Department of Psychology, University of Warwick, Coventry, CV4 7AL, United Kingdom*
[b]*School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom*
[c]*Department of English, Universität Wien, Campus d. Universität Wien, Spitalgasse 2-4/Hof 8.3, 1090 Wien, Austria*

Abstract

Linguistic variation is constrained by grammatical and social context, making the occurrence of particular variants at least somewhat predictable. We explore accommodation during interaction as a potential mechanism to explain this phenomenon. Specifically, we test a hypothesis derived from historical linguistics that interaction between categorical and variable users is inherently asymmetric: while variable users accommodate to their partners, categorical users are reluctant to do so, because it would mean violating the rules of their grammar. We ran two experiments in which participants learnt a miniature language featuring a variable or categorical grammatical marker and then used it to communicate. Our results support the asymmetric accommodation hypothesis: variably-trained participants accommodated to their categorically-trained partners, who tended not to change their behaviour during interaction. These results may reflect general social cognitive constraints on acquiring and using variable linguistic devices, and give insights into how small-scale interactive mechanisms may influence population-level linguistic phenomena.

Key words: artificial language; regularisation; unpredictable variation; interaction; alignment; accommodation

Email address: o.feher@warwick.ac.uk (Olga Fehér)

Introduction

Languages exhibit variation at all levels of organisation, but this variation is limited by grammar and social context. The ways in which linguistic units can be used reflect physiological, cognitive, socio-psychological, or functional constraints on language learning and verbal communication. A growing body of experimental work shows how language learning, use, and transmission (re-)shape patterns of linguistic variation. Here we explore how language-internal factors influence the ways in which languages are reshaped during language use. Our experiments are inspired by the phenomenon of *obligatorification* in language change, i.e. the tendency for constituents to shift from occurring variably and being pragmatically conditioned to being obligatory and grammatically conditioned. To provide a possible account for this tendency, we introduce the hypothesis of grammar-based asymmetric accommodation: when users of categorical and variable grammars interact, the latter will tend to accommodate to the former rather than vice versa, so that they will converge on categorical language use. We test this hypothesis experimentally, using artificial language learning and interaction paradigms, and find evidence consistent with grammar-based asymmetric accommodation. The paper thus introduces a new paradigm for testing mechanistic accounts of language change, and contributes to the growing literature seeking to explain fundamental properties of human language in terms of constraints operating on language learning and language use.

Learning, use, and the evolution of variation

Constraints on variation in natural language

Variation is an inherent property of natural languages. It occurs both synchronically, in the phonetic, morphological and syntactic choices speakers make when constructing utterances, and diachronically, as languages change over time. Nonetheless, it is tightly constrained: variants tend to be conditioned either on grammatical or on socio-pragmatic context (Givón, 1985).

Some variation is entirely deterministic. The English first person pronoun, for example, takes the form *I*, when it functions as a subject (as in *<u>I</u> like tennis*), and the form *me*, when it functions as an object (as in *He likes <u>me</u>* or *Give this to <u>me</u>*). The forms of German articles are determined by the (grammatical) gender of the nouns they determine: 'the man' is *<u>der</u> Mann*, 'the woman' is *<u>die</u> Frau*, and 'the car' is *<u>das</u> Auto*. When the choice of a constituent variant is conditioned by (one or more) other constituents in the linguistic signal, one speaks of morpho-syntactic, or grammatical conditioning. Such conditioning results in 'grammatical patterning' (Hockett, 1963), which is one of the definitional features of human language.

Deterministic conditioning is not necessarily grammatical however. It can also be pragmatic. For example, in many languages, including English, the use of count nouns in the singular requires the marking of reference relations by means of either definite or indefinite determinatives. The choice between the definite and indefinite is determined by the speaker's inferences about their addressees, specifically what the speaker thinks they know about the relevant utterance context: when they assume that a noun's unique referent is known, they choose *the*, otherwise they choose *a*.

Variation can also be probabilistic rather than deterministic. For instance, the so-called dative alternation in English (*I gave Jessie an apple* vs. *I gave an apple to Jessie*) is probabilistically conditioned on such parameters as the relative novelty of the referents of the two noun phrases, or their relative syntactic weight. Sociolinguistic variation can also be probabilistic: for instance, the pronunciation of English *-ing* (as in *finding, running*) takes one of two forms: [ɪŋ] or [ɪn], and speakers' choice varies according to the formality of the situation, the speaker's gender (Fischer, 1958), or their social status (Shuy et al., 1967).

In sum, natural linguistic variation tends not to be unpredictable or random. Instead, it is systematically constrained. Although conditioning factors may be complex and difficult to identify (Lass, 1984; Dixon, 1972; Labov, 1963), truly unpredictable, unconditioned, or 'free' variation seems to be rare.

The role of learning in constraining variation

What are the mechanisms that constrain variation in natural languages? Several converging lines of evidence suggest that biases in language acquisition play a crucial role. When adults learn new languages, they often use grammatical variants inconsistently (Newport, 1990; Johnson et al., 1996). Although the variants they produce may be conditioned by a range of factors, (Wolfram, 1985; Bayley, 1996), these factors work differently and idiosyncratically in different individuals. Thus, variability in the speech of adult learners is generally much higher than among native speakers. However, when children of adult second language learners are exposed to the variable and inconsistent output of their parents, they often eliminate the inconsistencies and regularise the language. Singleton & Newport (2004) describe the case of a deaf child who acquired American Sign Language from his hearing parents, both of whom had learnt it (imperfectly) as adults. Although the parents' signing contained highly variable and inconsistent morphology, the sign language of the child exhibited regular, consistent morphology.

A similar process is observed in creolisation: an example of new language formation that occurs when adults with different linguistic backgrounds are brought together and are under pressure to communicate (see DeGraff, 1999, for a review on creolization and language change). The pidgins (or early creole languages) which emerge in this situation tend to be highly variable, due to the diversity of grammatical structures of the contributing languages (e.g. Bickerton & Givón, 1976). Transmission of pidgins across speakers leads to the emergence of stable creole languages that exhibit grammatical properties characteristic of natural languages, such as reduced and grammatically conditioned variation. Some attribute these changes to child learners (Bickerton, 1981, 1984), while others argue for the important role of adult learners (Aitchison, 1996). For a review on regularization and creolization see Hudson Kam & Newport (2005).

Observational work is supported by experiments using artificial language paradigms. In these experiments, participants are exposed to a miniature, experimenter-designed language containing unpredictable variation and are then asked to reproduce that language. Artificial language paradigms have a long history as a tool for exploring statistical or distributional learning. They have been used extensively to study word segmentation (e.g. Saffran et al., 1996), word learning (e.g. Yu & Smith, 2007; Smith & Yu, 2008), the learning of grammatical categories (Frigo & McDonald, 1998; Gerken et al., 2005), and the acquisition of phonology (Chambers et al., 2010) and syntax (Reeder et al., 2013; Wonnacott et al., 2008, 2012) in both adults and children. A major advantage of artificial language paradigms is that they provide experimental control over learners' linguistic input (Aslin et al., 1998), allowing for the dissociation of age and linguistic experience. There is also evidence that artificial languages are processed similarly to natural languages by learners (Wonnacott et al., 2008; Magnuson et al., 2003; Ettlinger et al., 2016; Fehér et al., 2016).

These paradigms have been used to explore how learning biases shape language, for example when learners acquire a language with synonymous forms whose use varies unpredictably (unlike in a natural language). Pioneering experiments demonstrated that children eliminate unpredictable variation during learning, by eliminating all but one of the competing forms (Hudson

Kam & Newport, 2005, 2009) - just as observed by Singleton & Newport (2004) in a natural language setting. While adult learners are more likely to reproduce the probabilistic usage of variants and match the statistics of their input (known as probability matching), adults also eliminate variability when that variability is complex (Hudson Kam & Newport, 2005, 2009; Hudson Kam, 2009) or when they have reason to believe that the variation is random rather than systematic (Perfors, 2016). On the other hand, children's preferences for regularity are reduced if the learning task is simplified, e.g. by mixing novel function words and grammatical structures with familiar English vocabulary (Wonnacott, 2011).

Related work explores how biases in learning can accumulate to shape languages over longer time-spans. In experiments by Reali & Griffiths (2009), Smith & Wonnacott (2010), Smith et al. (2017) and Vihman et al. (2018), an artificial language exhibiting unpredictable variation is transmitted across chains of adult learners in iterated learning experiments, where the language produced by one learner becomes the target language for the next learner in a transmission chain. In these experiments, participants gradually eliminate unpredictability, thereby revealing cumulative effects of weak individual-level biases: while no single individual reshapes the language radically, each individual in the chain increases its regularity subtly. When such small changes accumulate, they eventually produce highly regular systems where variation is either eliminated entirely (Reali & Griffiths, 2009) or is preserved but becomes grammatically conditioned (Smith & Wonnacott, 2010; Smith et al., 2017; Vihman et al., 2018). This finding is in line with a growing body of experimental work showing how universal structural properties of language emerge from learning biases when learning processes are iterated (see e.g. Kirby et al., 2014, for review).

The role of language use in constraining variation

Another important mechanism that shapes language structure is communicative interaction (cf. e.g. Bybee & Beckner, 2009; Ibbotson, 2013; Lieven, 2014; Lieven & Tomasello, 2008; Tomasello, 2003). Speakers acquire and use language interactively in a rich social environment. They learn not only by observation, but also by interacting with other language users and observing such interactions, and interaction can therefore shape linguistic systems. When speakers adapt their language use to meet their communicative needs, this can result in innovation and can change the linguistic conventions of a community (e.g. Heine, 1997; Croft, 2000). For example, when linguistic forms occur frequently, their occurrence becomes more predictable, and speakers can afford to pronounce them less distinctively. This may affect their mental representations, and may ultimately change the structure of a language (e.g. Bybee, 2001, 2006; Wedel, 2007; Garrod & Pickering, 2013).

To become conventionalised in a language, of course, innovative uses need to spread in a community, and one way in which this can happen is through a process known by the name of either *accommodation* (Coupland, 2010) or *alignment* (Pickering & Garrod, 2004). Both labels refer to the phenomenon of interlocutors modifying their speech to match that of their partners during communicative interaction; the two distinct terms reflect two different approaches, highlighting different aspects of communication as the major driving force behind the observed convergence. *Accommodation theory* emphasises the influence of social factors (Coupland, 1984; Soliz & Giles, 2014; Giles, 1984; Giles et al., 1991; Giles & Ogay, 2007; Trudgill, 2008), but it also acknowledges the importance of language-internal features, particularly their perceptual salience. For instance, when English and American speakers interact, the post-vocalic /r/[1] in the speech of the latter is easy to perceive and therefore likely to be emulated (MacLeod, 2012). *Alignment-based accounts* on the other hand stress the automaticity of convergence. According to Pickering & Garrod (2004), convergence is caused by a simple priming mechanism: hearers activate the linguistic representations of the forms they perceive and this makes them more likely to use the same forms when they speak. Priming occurs at various levels of linguistic representation: phonetic (Giles et al., 1991), lexical (Brennan & Clark, 1996; Garrod & Anderson, 1987), semantic (Garrod & Anderson, 1987; Garrod & Clark, 1993), and structural (Bock, 1986; Gries, 2005). Research on structural priming has demonstrated that priming rates are influenced by people's beliefs about their interlocutors (including e.g. beliefs about their linguistic knowledge: Branigan et al., 2011; Loy & Smith, submitted).

Several recent experimental studies have shed light on how processes occurring in communication might restructure unpredictably variable aspects of a language. Perfors (2016) found that participants trained on a variable input language produced more regular output when instructed to use the language as they think other participants might use it (in the absence of actual

[1] More precisely: /r/ that does not occur before a vowel, i.e. /r/ in words such as *car* or *cart*.

communication). Similarly, Fehér et al. (2016) found that variation was reduced during communicative interaction. This tendency to reduce variation during interaction could reflect active reasoning about the communicative consequences of variation. Deviations from a conventional way of conveying a particular idea can easily be taken to signal a difference in meaning (e.g. Horn, 1984; Clark, 1988). Therefore, producing unpredictable linguistic variation during communication might be dysfunctional: confronted with unpredictably alternating variants of a form, listeners might erroneously infer that the variation is meaningful after all (i.e. that each variant expresses something slightly different).

A hypothesis: grammar-based asymmetric accommodation

In the interaction-based experiment reported in Fehér et al. (2016), participants were trained on a shared target language that exhibited variation. Prior to interaction, participants typically reproduced the variable nature of their input successfully; during interaction, they converged with their partners in the way they used the language, eliminating variation. Here we extend this work to explore how this process of convergence unfolds when pairs of participants are trained on languages which differ systematically and qualitatively. In particular, we explore situations (motivated by cases of obligatorification in language change, discussed below) where one member of an interacting pair is trained on data that suggest categorical use of a given variant, whereas their interlocutor sees that variant occurring probabilistically.

The hypothesis we test is that the difference between categorical and probabilistic conditioning of linguistic constituents biases the direction of accommodation in favour of the former. In other words, we hypothesise that speakers who make variable use of a constituent will find it easier to accommodate to speakers who use the same constituent categorically in specific grammatical contexts. This strikes us as plausible, because all variable users need to do in order to emulate categorical usage is to make maximal use of an option they already have in their grammar. On the other hand, in order for categorical users to accommodate successfully to their variable interlocutor, they would not only have to violate a constraint in their grammar, but also uncover the (potentially subtle) conditions that govern their partner's choices. Since in such cases the direction of accommodation would not reflect social (power) relations between the participants, but would be primarily determined by differences between the grammars of the interlocutors, we dub our hypothesis *grammar-based asymmetric accommodation*.

An example from the history of English

Our hypothesis receives support from the histories of natural languages, which provide rich evidence of changes where optional variants become obligatory in specific grammatical contexts. An example of such a change is the development of optionally used demonstrative pronouns into articles that are obligatory in certain noun phrases. Although this change has occurred in many languages (see e.g. Himmelmann, 1997; van de Velde, 2010; Vincent, 1997), we briefly describe the emergence of definite articles in late Old English to illustrate it (for details see Sommerer, 2011, 2012, and the references therein).

The English article *the* derives from the masculine nominative singular *se* of the Old English deictic demonstrative *se – seo – þæt*. A defining feature that distinguishes articles from demonstratives is that they are grammatically obligatory under certain conditions. Thus, the English definite article must be used whenever a noun phrase headed by a common count noun refers to a unique entity (or set of entities) identified by the interlocutors. Its demonstrative predecessor, on the other hand, was used only optionally in such contexts. For example, it is present in the Old English example (1) below, but not in (2) or (3).

(1) þa Eadmund clypode ænne bisceop [...] þa forhtode
Then Edmund summoned a bishop [...] then was frightened
se bisceop
the bishop

'Then Edmund summoned a bishop [...] **the bishop** was frightened.'
(Ælfric *Saints* XXXII.56)

(2) Stonc ða æfter ∅ **stane** stearcheort onfand ∅ **feondes**
jumped then behind ∅ **rock** stouthearted, found ∅ **enemy's**
fotlast
footprint

'He jumped behind **the/a rock**, courageously, and discovered **the enemy's** footprint.'
(*Beowulf* 2288)

(3) Gecyste þa ∅ **cyning** ∅ **æþelum** god, ∅ **þeoden** ∅
kissed then ∅ **king** ∅ **nobles**.DAT good ∅ **Lord** of ∅
Scyldinga, ∅ **ðegn** betstan
Scyldings ∅ **warrior** best

'**The good king** of **the nobles**, **the lord** of **the Shieldings**, kissed **the best warrior**.'
(*Beowulf* 1870)

What is important in this case of article emergence is that a constituent whose use had been pragmatically and probabilistically conditioned became grammatically obligatory. Thus, the Old English demonstrative was used for indicating that a noun phrase had a unique referent, but it was used only optionally, i.e. when speakers believed that it was helpful or even necessary to indicate this. In cases where the referent of a noun phrase was evident, there was no need for an explicit marker. Of course, assessing whether an explicit reference marker should be used or not would have depended on a variety of situational and social factors. On the one hand, for instance, speakers would have to estimate what their addressees could be expected to know and be aware of, and on the other, they would have to decide how polite and communicatively helpful they should be. Such assessments are highly subjective and may reflect variable, culture-specific politeness conventions (see e.g. Leafgren, 2002; Leech, 2014). Therefore, demonstrative use would have been probabilistically rather than categorically conditioned.

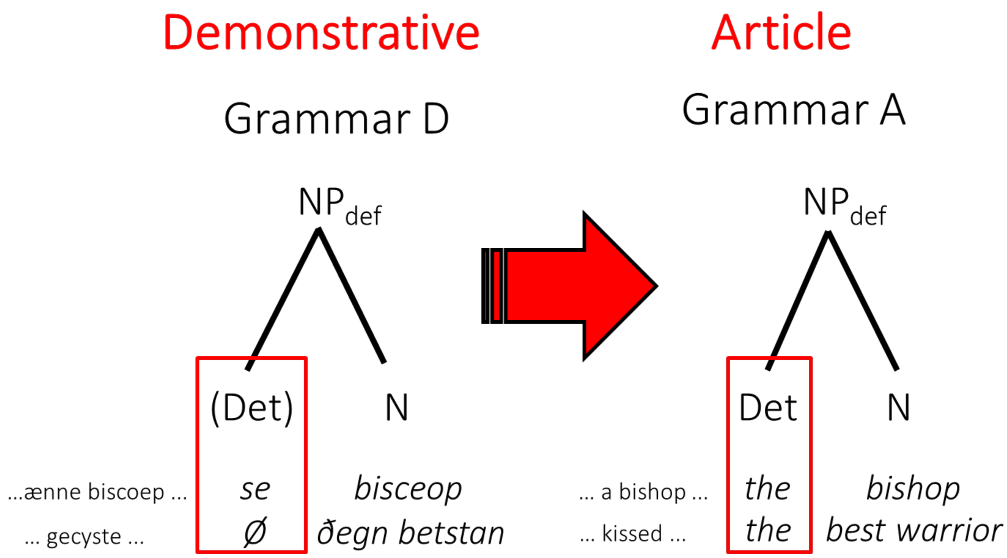

Figure 1: The change from demonstrative to definite article in Old English.

In contrast, the newly emergent article had to be used whenever a noun phrase had a referent that was assumed to be known to both interlocutors, no matter if the identity of that referent was self-evident or whether the article was required to facilitate its identification. Thus, a crucial difference between the demonstrative and newly emerging article was that the former was still used variably and was pragmatically conditioned, while the latter was obligatory and grammatically conditioned, as shown in Fig. 1.

Obligatorification as a general process

Processes by which the (pragmatic) probabilistic conditioning of a constituent comes to be categorical and grammatical are attested not only in article emergence. They occur frequently in changes known collectively as *grammaticalisation*. Another case from the history of English would be the development of *do* into an obligatory maker of questions and negations, and the literature provides many examples from other languages as well (see e.g. Diewald & Smirnova, 2010; Reinöhl, 2016). In studies of grammaticalisation, the establishment of categorical grammatical conditioning is called *obligatorification*. In obligatorification a linguistic sign loses "paradigmatic [... and] syntagmatic variability[, i.e.] the possibility of using other signs in its stead or of omitting it altogether[, and ...] the possibility of shifting it around in its construction" (Lehmann, 1985).

Although instances of obligatorification are widely attested in the histories of languages, the focus of historical linguistic research has been mostly on identifying and describing relevant cases. As to their explanation, the roles of usage and cognition in grammaticalisation have been studied intensely, but neither the potential role of interaction nor the specific aspect of obligatorification have received much attention. An explicitly cognitive theory of grammaticalisation is represented in the work of Joan Bybee (e. g. Bybee, 2010), for example. There, the emergence of obligatory constituent use is conceived of as a gradual process, in which the productivity of grammatical patterns gets extended and maximised. Frequency and analogy are shown to play important roles, but interaction and accommodation are not specifically considered. Therefore, our study complements extant work on grammaticalisation in that respect.

Our focus is on the role of interaction in spreading obligatory usage patterns in communities, and our conceptual starting point is a mixed community of speakers, where some use a variant categorically in specific grammatical contexts, while others use it in the same contexts but variably so. Several viable hypotheses for how such scenarios may arise in the first place can be derived from the literature. For instance, (over-)generalisation during language acquisition (Wolff, 1982) would represent a plausible mechanism. Young children are more likely than adults to regularise probabilistic input (Hudson Kam & Newport, 2005, 2009), and this regularisation can involve over-using the most frequent form in the input. In the case of article emergence, for example, a child who is exposed to input in which a sufficiently large proportion of noun phrases with definite reference take the article might infer that the article is to be used in all of them. At the same time, maximal article usage would not be perceived as illicit by adult speakers, whose grammar provides the option after all. Thus, it may come to stabilise in the learner's language. In other words, it does not strike us as implausible that categorical use of a constituent should emerge in some individuals in a community where it is used optionally, albeit frequently. At the same time, and as pointed out above, this is not the issue our paper addresses, and will require more research in its own right.

Instead, we ask whether optional or categorical usage patterns are more likely to be adopted through accommodation in communicative interaction, and hypothesise that the latter is the case. As indicated above, we suspect that the categorical, grammatically conditioned use of a constituent should be easy to emulate by speakers who have learned to use a constituent optionally under specific pragmatic conditions. In contrast, speakers who have learnt to use it categorically in specific grammatical contexts will find it difficult to violate their grammar and to imitate patterns that are probabilistically variable. Should this be the case, it would predict that categorical and variable users will converge on categorical use when they accommodate to each other. This would predict, in turn, that categorical usage patterns that emerge in a speech community will spread at the cost of variable ones, which would serve to explain the frequency of obligatorification in language change.

This study

In order to test what we have called the grammar-based asymmetric accommodation hypothesis, we use experimental techniques that have been developed for studying the acquisition and use of variable linguistic systems (reviewed above). The specific experiments reported here were designed to test whether and under what conditions interaction leads to obligatorification. Although evidence of obligatorification comes from language history, our experiments do not attempt to replicate a particular language change (such as the emergence of articles in English). Instead, we employ a specifically designed artificial language to address the problem in the most general terms possible.

In Experiment 1 we test whether interaction results in convergence between variably-trained interlocutors and in a loss of variation overall, even in situations where individuals differ markedly in their pre-interaction use of a variable grammatical marker. Experiment 1 also provides a control condition for Experiment 2, where we directly test the grammar-based asymmetric accommodation hypothesis: do we see an asymmetry in accommodation between interlocutors, such that individuals with variable grammars accommodate to categorical users but not vice versa?

Although these experiments were inspired by the emergence of the English article, we simplify away from the details of this case in two respects. First, we test number marking instead of definiteness. This is because number distinctions can be easily represented and controlled in experimental setups, whereas distinctions between referents that participants want to count as either having been identified or not depend so strongly on their subjective interpretations that they cannot be reliably controlled in experiments.

Second, when we train participants on variable use, we expose them to random variation rather than to variation that is subtly conditioned by the complex interplay of various pragmatic factors (such as assumptions about shared knowledge and politeness). The rationale behind this simplification is twofold. On the one hand, participants trained on variable use may in any case apply their own hypotheses about potential conditioning factors when trying to reproduce variation. On the other hand, categorically trained participants are unlikely to be able to distinguish between random variation and complexly conditioned variation when they are exposed to it.

Experiment 1

In Experiment 1 we test whether interaction results in convergence between variably-trained interlocutors and in a loss of variation overall, even in situations where individuals differ markedly in their pre-interaction use of a variable grammatical marker.

Method

Participants

Eighty participants were recruited from the University of Edinburgh's Student and Graduate Employment service and the University of Warwick's sign-up system for Psychology and Behavioural Science research. Participants were recruited to take part in a miniature language communication experiment and were paid £8-10 for their participation (depending on the time it took them to finish the experiment).[2]

[2] We initially ran 40 participants for Experiment 1, and 40 for Experiment 2, all tested at the University of Edinburgh. During the review process we were asked to increase our sample size, and therefore doubled the sample size in both experiments, with the second batch of participants recruited at the University of Warwick. In general the pattern of results in the data collected across the two sites are highly consistent in both experiments, and there were no cases where an effect which was significant in the original data set (i.e. $p < .05$) fell below this significance threshold in the enlarged dataset; there are a small number of cases where effects which were marginal in the original dataset are now significant, or where effects were n.s. in the original dataset but are now marginal. Since these cases do not change our interpretation of our results we do not flag them up here. The full dataset, including an indication of the testing site for each participant, is available online, link provided at the start of the Results section.

Procedure: summary

Participants used an online system to sign up for the experiment individually, but were scheduled to arrive in the lab in pairs. After briefing, they were seated in isolation in sound-proof booths, and worked through a computer program which presented and tested them on an artificial language, and then allowed them to use that language to communicate remotely with their partner, another participant going through the experiment at the same time. The language was text-based: participants observed pictures and text displayed on the screen and entered their responses using the keyboard.

Procedure: Language Training and Testing

Participants progressed through a six-stage training and testing regime. *1) Noun training:* Participants viewed pictures of six cartoon animals (bird, elephant, frog, insect, pig, shark) along with nonsense nouns which were intended to be memorable and transparently related to their associated referent animal (*beeko*, *trunko*, *hoppo*, *bugo*, *oinko* and *fino*). Each presentation lasted 3 seconds, after which the text (but not the picture) disappeared and participants were instructed to retype that text. Participants received 4 blocks of training, each consisting of one presentation of each noun in random order. Presentation order for the two members of a pair was randomised independently throughout training and individual testing. In order to keep the participants roughly synchronised, participants were only allowed to progress to the next block of training/testing when their partner was also ready to begin the corresponding block.

2) Noun testing: Participants were presented with a picture of an animal, without accompanying text, and were asked to provide the appropriate label. Participants were tested on each animal once, in random order.

3) Sentence training: Participants were exposed to sentences paired with visual scenes. Scenes showed either single animals or pairs of animals (of the same type) performing one of two possible actions, depicted graphically using arrows: either a straight left-to-right movement, or a bouncing left-to-right movement. Sentences were presented in the same manner as nouns (participants viewed a scene plus text, then retyped the text). The language is presented in Fig. 2: each description consisted of a nonsense verb (*wooshla* for straight movement, *boingla* for bouncing movement), a noun (the same nouns as in noun training) and a number marker. Each pair of participants was assigned two number markers, one which was used to mark the singular and one which was used to mark the plural, selected randomly from the set

$S_{Singular} \rightarrow V\ N\ M_1\ \ (p = \alpha)$
$S_{Singular} \rightarrow V\ N\ \ (p = 1 - \alpha)$
$S_{Plural} \rightarrow V\ N\ M_2$

$N \rightarrow$ {beeko, trunko, hoppo, bugo, oinko, fino}
$V \rightarrow$ {wooshla, boingla}
$M_1 \rightarrow$ bup
$M_2 \rightarrow$ dak

Figure 2: The grammar of the target language. The language explicitly marks the plural with a marker M_2 (randomly pre-selected from a list of 8 possible markers — in the example grammar, the plural marker is *dak*), but the singular is either marked with M_1 (selected from the same list of possible markers — in this example, the overt singular marker is *bup*) or left unmarked. The probability, α, with which the singular marker appears varies according to condition; the possible values of α in Experiment 1 are 1/6, 1/3, 2/3 or 5/6.

bup, dak, jeb, kem, pag, tid, wib, yav. For instance, if the randomly-selected markers were *bup* and *dak*, then one bird moving straight would be labelled *wooshla beeko bup* or *wooshla beeko* (depending on whether the singular was marked, see below), and two sharks bouncing would be labelled *boingla fino dak.* Each of the 24 possible scenes (6 animals x 2 motions x 2 numbers) was presented six times during training (in six blocks, order randomised within blocks).
4) Recall test 1: Participants viewed the same 24 scenes without accompanying text and were asked to enter the appropriate sentence. Each of the 24 scenes was presented three times (in three blocks, order randomised within blocks).
5) Interactive testing: Participants played a director-matcher game in which they alternated describing a scene for their partner, and selecting a scene based on their partner's description. When directing, participants were presented with a scene (drawn from the set of 24 possible scenes) and prompted to type the description so their partner could identify it. This description was then passed to their partner[3], who had to identify the correct scene (by

[3]In fact the closest legal description was passed to their partner, to prevent participants communicating using English or any system other than the language they were trained on: the string produced by the director was checked against all 36 legal strings in the language the participants were trained on (2 verbs x 6 nouns x three possible markers [null, M_1,

button-press) from an array of 8 possibilities: these 8 possibilities contained two animal types (the animal in the director's scene plus one other randomly-selected animal type), both motions (straight and bounce) and both numbers (singular and plural), and thus were guaranteed to contain the target but in themselves provide no information as to the correct target. After each trial both participants then received feedback (either success or failure) and an updated score ("Score so far: X out of Y"). Participants played 96 such communication games, organised into two blocks of 48 trials, such that each participant directed once for each possible scene within each block (order randomised within blocks, a randomly-selected member of the pair directing first in each block and the participants alternating roles for the remainder of the block).

6) Recall test 2: As in recall test 1, participants once again viewed the same 24 scenes without accompanying text and were asked to enter the appropriate sentence. Participants were specifically instructed to remember the language they were initially taught. Each of the 24 scenes was presented three times (in three blocks, order randomised within blocks). By comparing this second post-interaction recall test to pre-interaction recall we can evaluate whether any changes in marker use occurring during interaction persist beyond that interaction.

Manipulation: Variable marking of the singular

The training language provided post-nominal particles to mark singular and plural (Fig. 2). The plural was consistently marked for all participants throughout training: every sentence labelling a scene featuring two animals included the appropriate post-nominal marker. We manipulated the frequency with which participants saw overt marking of the singular during training: participants saw singular marking on 5 in 6 singulars (for convenience, we refer to this as 83% marking) with the remainder unmarked (i.e. in unmarked sentences, the sentence contained only the verb and the noun),

M_2]), and the closest legal string (by Levenstein string-edit distance) was transferred to the matcher. This is purely an issue of experimental control: our intention was to constrain the effects of interaction to altering the frequencies of the linguistic variants provided in the target language, rather than allowing participants to introduce new variants and depart radically from the target language during interaction. This substantially simplifies our analysis, but also constrains the solutions participants arrive at to those which speak directly to the hypotheses this experiment was designed to test. In practice, errors were rare and essentially restricted to typos (e.g. *beko* instead of *beeko*).

or 2 in 3 singulars marked (66% marking), or 1 in 3 singulars marked (33% marking), or 1 in 6 singulars marked (17% marking). The training data was constructed such that singular marking was unconditioned and unpredictable: across the 6 blocks of training, every noun was marked for singular an equal number of times, and every verb appeared with a marked singular an equal number of times.

Participants within a pair differed in the language they were trained on. We ran two combinations of pairings. We will refer to the participant trained on the higher frequency of singular marking as *P1* and the participant trained on the lower frequency as *P2*. In the 66-33 condition (20 participant pairs), P1 was trained on 66% marking, P2 was trained on 33% marking; in the 83-17 condition (20 participant pairs), P1 was trained on 83% marking, P2 on 17% marking. These two conditions allow us to test whether interaction leads to the reduction or elimination of unpredictable variation in singular marking, and whether this is dependent on the degree of similarity between participants prior to interacting: the difference in frequency of marked singulars during training is much greater in the 83-17 condition than the 66-33 condition.

Analyses

Each participant produced 192 typed descriptions across the three test phases of the experiment: 72 at recall test 1 (henceforth Recall 1), 48 during interaction, 72 at recall test 2 (Recall 2). Our hypotheses concern the marking of the singular, which is marked variably during training. For the purposes of statistical analysis, we therefore automatically coded each description which referred to a scene in which there was a single animal in the following way. Taking the description typed by the participant, we split that description into a series of words, by splitting the string at spaces (ignoring leading or trailing whitespace). Those words were then categorised as Noun, Verb or Marker, by comparison to the list of 16 legal words (6 nouns, 2 verbs, 8 possible number markers), by identifying the closest legal word (by Levenstein distance) — for instance, *beko* would be classified as a Noun, as its closest legal match (*beeko*) is a Noun. This process generates a list of categories for each typed description. Descriptions consisting of the sequence Verb-Noun were classified as unmarked singulars; descriptions consisting of the sequence Verb-Noun-Marker were classed as marked singulars; all other sequences of categories were classed as NA, and excluded from the analyses that follow.

This produces a binary dependent variable for every trial, which makes this data in principle suitable for analysis using logistic regression. However, the nature of the data (many participants produce marked or unmarked singulars categorically during interaction, particularly in Experiment 2) leads to extensive problems with convergence when using e.g. glmer in R (Bates et al., 2015). We therefore calculated the proportion of trials for each participant which feature a marked singular at a given phase of the experiment. The resulting distributions of proportions are highly non-normal; we therefore exclusively use non-parametric inferential statistics. To evaluate the degree of change we calculated by-participant differences (e.g. difference between the training proportion of marked singulars and that produced at Recall 1; difference in proportion of marked singulars produced at Recall 1 and during interaction) and then run statistics on those difference scores. We use the Wilcoxon signed rank test (testing whether the median difference score is significantly different from 0, i.e. do participants change?). We use the Wilcoxon rank sum test for comparisons between groups (e.g. does the amount of change seen in P1s differ from that seen in P2s?; does the amount of change in the 66-33 condition differ from that seen in the 83-17 condition?). In order to test for statistical interactions in between-group factors (i.e. do P1 and P2 differ between conditions in the extent to which they change their behaviour?) we calculate a difference-in-change score for each pair (change in marked singular use for P1 minus change in marked singular use for P2) and then compare those difference-in-change scores across conditions using the Wilcoxon rank sum test: a significant difference indicates an interaction, i.e. the extent to which P1 and P2 differ depends on condition. Finally, we also analyse changes in within-pair difference in marker use at various phases of the experiment, i.e. do interacting pairs become more similar in their use of the singular marker during interaction? To do this we calculate a within-pair difference in marker use, which is simply the absolute difference in marker use between P1 and P2 in a given pair, and then look at changes in those within-pair difference scores over various phases of the experiment as above. Both the rank sum and signed rank test statistics are computed using the `wilcox.test` command in R version 3.5.0 (R Core Team, 2018): in R the rank sum test returns a test statistic W, the signed rank test returns a test statistic V.

The full dataset and all analysis code, as well as various supplemental figures, for this experiment and Experiment 2 are available online at `https:`

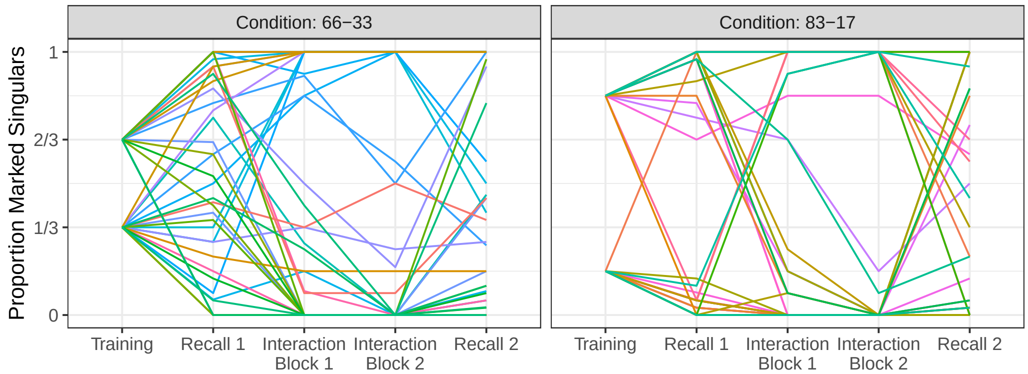

Figure 3: Proportion of trials in which the singular was marked, in training (determined by condition), Recall 1, interaction (split by block) and the post-interaction Recall 2. Each pair is represented by two lines, one per participant, sharing the same colour: alignment between participants is therefore reflected in lines of matching colour converging. See also Fig. A.1

`//github.com/kennysmithed/Asymmetric`.[4]

Results

Performance during the communicative portion of the task was extremely high throughout, and varied little across conditions or across the two blocks in interaction: the mean number of successful trials (in which the matcher selected the picture presented to the director) was 46.625 out of 48 in the 66-33 condition (46.65 in the first block of interaction, 46.6 in the second), 46.9 in the 83-17 condition (46.55 in block 1; 47.25 in block 2).

Our main dependent variable of interest is participants' use of the singular marker. Fig. 3 shows the full data for use of the singular marker across training, individual testing and two blocks of interaction (see Fig. A.1 for separate by-pair plots). Fig. 4 provides means for the various phases.

In both conditions, we see variable responses during Recall 1, and rapid alignment during interaction. Most pairs align on either systematic use (11 pairs) or systematic non-use (20 pairs), with an overall preference for non-

[4]Note for review: we will archive the data on the University of Edinburgh's DataShare service (which provides curation and long-term archive support) on acceptance; we are using github in the interim.

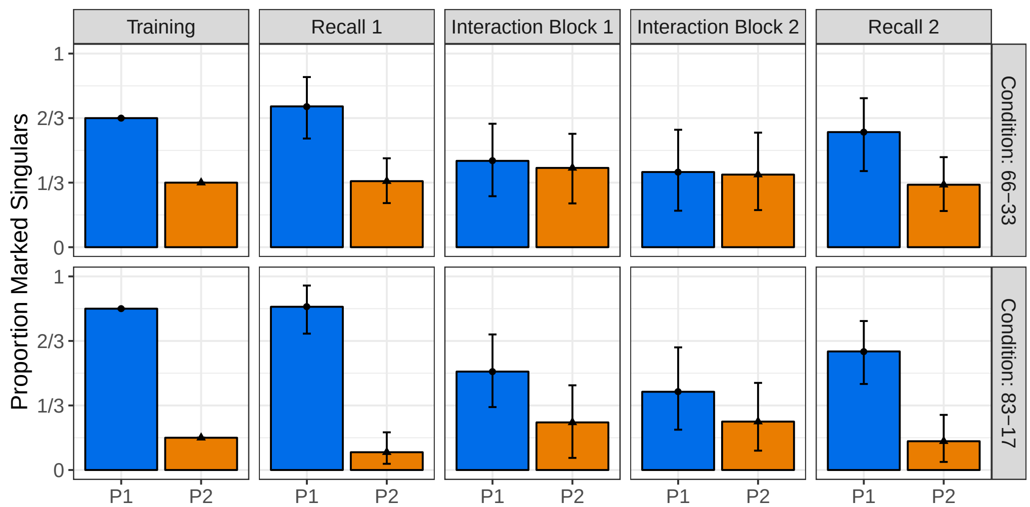

Figure 4: Mean proportion of trials in which the singular was marked in training (determined by condition), Recall 1, interaction (split by block) and the post-interaction Recall 2, for the 66-33 condition (upper panels) and 83-17 condition (lower panels). Error bars indicate bootstrapped 95% confidence intervals, obtained using 10,000 bootstrap samples and the percentile method. Note that these error bars reflect the variance within each participant group at each stage, and cannot be interpreted as within-subjects confidence intervals indicating reliability of change within subjects.

use reflected in the low average frequency of marking of singulars during interaction. Finally, some but not all participants return to being variable users in the post-interaction Recall 2.

The statistical analyses in the following sections seek to answer four questions. Firstly, did participants *probability match* during individual testing, i.e. reproduce the marker frequency they were trained on? Secondly, did participants change their use of the singular marker during interaction, relative to their use of the marker during Recall 1? Third, did participants *align* during interaction, i.e. come to use the singular marker in the same way as their partner, and if so, was this modulated by similarity of their training data, i.e. did it differ across conditions? Fourth, did the effects of interaction persist into the post-interaction recall test — i.e., did participants revert to their pre-interaction recollection of the language, or was their estimate of the frequency of singular marking changed by interaction? We evaluate these questions using two measures: we measure how the participants' use of the singular marker changes across the course of the experiment (see Figure 5), and how within-pair difference (i.e. the absolute difference between the proportion of marked singulars produced by P1 and P2, see Figure 6) changes across the course of the experiment.

Change in marker usage

Figure 5[5] plots the change in marker usage across three key phases of our experiment, comparing proportion of marked singulars produced during Recall 1 to that seen during training (upper figure); change from Recall 1 to block 2 of interaction (middle), and change from Recall 1 to Recall 2 (lower).

The change in frequency of singular marking between participants' training data and their productions in Recall 1 indexes the extent to which participants are probability matching: change values of around 0 are indicative of probability matching, i.e. reproducing the singular marker in the proportion seen during training. During Recall 1, participants exhibit a great

[5]Annotations associated with individual bars indicate significance of comparison to 0, i.e. whether the amount of change is significantly different from 0 (n.s. $= p > .1$; * $=$ $p < .05$; ** $= p < .01$; *** $= p < .001$); differences between conditions are indicated by horizontal bars and an associated annotation. The absence of an annotation indicates the specific test was not run — in particular, note that we do not test each condition separately unless licensed to do so by a significant difference between conditions or a significant interaction.

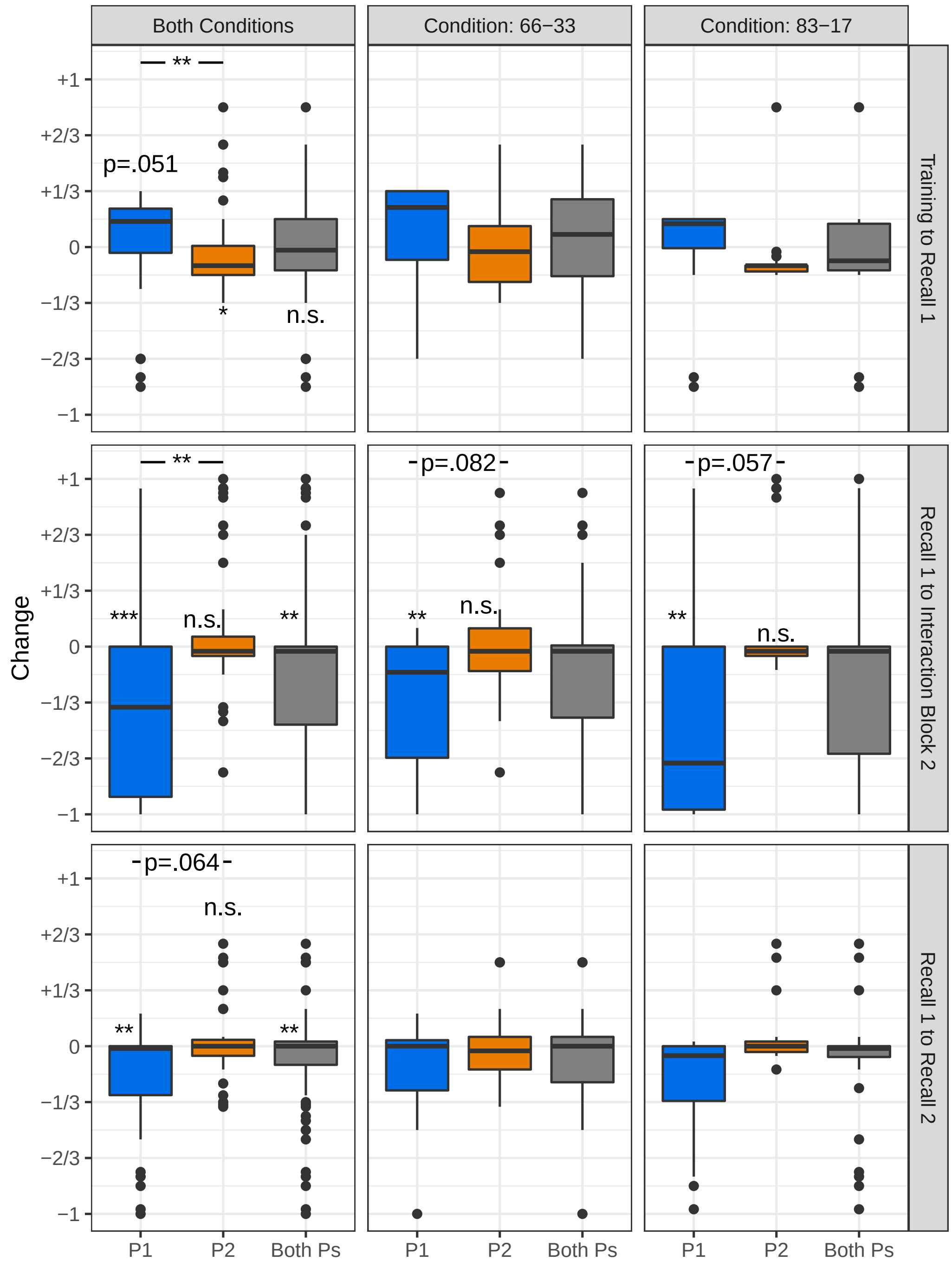

Figure 5: Change in proportion of marked singulars from training to Recall 1 (upper), from Recall 1 to block 2 of interaction (middle) and from Recall 1 to Recall 2 (lower). In all cases, change is calculated as proportion of marked singulars at the later stage of the experiment minus the proportion of marked singulars at the earlier stage - i.e. positive values indicate an increase in singular marking, negative values indicate a decrease. We show the data for both conditions combined, in addition to separated by condition, since the analyses in the main text often indicate no significant effects of condition and no condition by P1/P2 interaction. See footnote 5 for details of annotations. These are standard Tukey boxplots.

deal of variation in marker use, with some completely eliminating one of the markers (see Fig. A.1). The participant population collectively exhibit probability matching behaviour: collapsing across conditions and P1/P2, the change from training to Recall 1 is not significantly different from zero ($V = 1643$, $p = .611$); however, while there is no significant difference between conditions in the training-to-Recall 1 difference scores (n. s. effect of condition: $W = 964$, $p = .115$), there is a significant difference between P1 (trained on the higher proportion of marked singulars) and P2 ($W = 1099.5, p = .004$); the interaction between condition and P1/P2 is not significant ($W = 207, p = .860$), suggesting this difference between P1 and P2 is roughly equivalent in both conditions. Considering P1 and P2 data separately, and collapsing across condition, P1s mark singulars marginally more frequently than in their input (the change from training to Recall 1 is marginally significantly different from 0, $V = 530, p = .051$), while P2s produce fewer marked singulars than exemplified in their input (change is significantly less than 0, P2: $V = 240.5, p = .037$). This pattern of results suggest that participants are drawn somewhat towards the regular extremes of either always or never marking the singular, depending on whether the marked singular is the more or less frequent option in their input; a similar tendency is seen in other studies of variation learning, e.g. Ferdinand et al. (2019).

The change in frequency of singular marking between Recall 1 and interaction (specifically, the second block of interaction, allowing for the possibility that marker use is fluid during the early stages of interaction) allows us to test whether participants continue to reproduce similar amounts of variability during interaction, or whether interaction changes their use of the singular marker. These change values are shown in the middle panel of Figure 5. Interaction substantially changes marker use in both conditions (n. s. difference between conditions: $W = 816.5, p = .877$); however, P1 and P2 show different amounts of change (significant effect of the P1/P2 difference, collapsing across conditions: $W = 531, p = .009$), and there is some evidence of an interaction between condition and P1/P2 ($W = 277, p = .038$). The change in singular marking over the entire data set is significantly negative, indicating a tendency to reduce singular marking during interaction (the change from Recall 1 to interaction block 2 is significantly non-zero, $V = 548.5, p = .007$); considering P1 (trained on the higher proportion of marked singulars) and P2 (trained on the lower proportion) separately and collapsing across conditions, P1s show this reduction in singular marking

($V = 33, p < .001$) whereas P2s do not ($V = 283, p = .972$) as they and their partner converge on a system which was closer to their infrequent use of the singular marker during Recall 1. Indeed, as can be seen in Fig. 3, most pairs converge during interaction on systems which either never or (more rarely) always mark the singular. Given the indication of an interaction between condition and P1/P2, we also consider each condition separately; both conditions show marginal differences between P1s and P2s, although this difference is clearer in the 83-17 condition (66-33 condition: $W = 135.5, p = .082$; 83-17 condition: $W = 130, p = .057$); in both conditions P1s show a significant reduction in singular marking whereas P2s do not (66-33 condition, P1s: $V = 6, p = .006$; 83-17 condition, P1s: $V = 10, p = .008$; 66-33 condition, P2s: $V = 85, p = 1$; 83-17 condition, P2s: $V = 54, p = .754$).

Finally, the change in singular marking from Recall 1 (pre-interaction) to Recall 2 (post-interaction) indicates whether the reduction in singular marking during interaction persists beyond that interaction — in other words, during Recall 2, did participants revert to their pre-interaction recollection of the language, or was their recollection of the frequency of singular marking in their training changed by their behaviour and their partner's behaviour during interaction? The lower panel of Figure 5 shows this measure of the lasting effects of interaction. The difference between conditions is not significant ($W = 861.5, p = .554$); collapsing across conditions, there is a marginal difference between P1 and P2 ($W = 609, p = .064$) suggesting that the participants might differ in the extent to which interaction leads to lasting changes in singular marking; the absence of an interaction between condition and P1/P2 ($W = 247, p = .208$) suggests this P1/P2 difference is roughly equivalent across conditions. Collapsing across conditions and P1/P2, our entire data set shows a significantly non-zero change ($V = 542, p = .010$), suggesting that there is a small but measurable tendency for the reduction in singular marking during interaction to persist beyond the duration of the interaction. An analysis of P1 and P2 separately, collapsing across condition, suggests this effect is largely borne by the P1 participants, who were trained on more frequent singular marking and changed their behaviour more during interaction: P1s show a significant reduction in marker use from Recall 1 to Recall 2 ($V = 61.5, p = .001$), whereas P2s do not ($V = 229.5, p = .724$).

Change in within-pair differences

The results above are for individual participants, and do not speak directly to the hypothesis that interlocutors will converge in their use of the singular marker during interaction. As can be seen in Fig. 3, there is a strong tendency for pairs of participants to converge on a shared system of using the singular marker. Figure 6 plots within-pair difference in singular marking across the various stages of the experiment, as well as the change in within-pair difference at several key stages.

Within-pair differences during Recall 1 reflect the differences in the frequency of singular marking in the participants' training data, as expected given that our participants are probability matching or even pulling apart slightly as they move towards a more extreme use of the singular marker. However, within-pair differences sharply reduce during interaction, as is clear from the lower panel of Figure 6 showing change in within-pair difference from Recall 1 to interaction block 2. As suggested by the Figure, there is at most a marginal difference between conditions in the amount of change in within-pair difference ($W = 265.5$, $p = .079$); across the whole data set there is a significant reduction in within-pair difference from recall test 1 to interaction block 2, indicating convergence on a shared system of marker use ($V = 26.5, p < .001$), an effect which is robust in both conditions if considered separately (66-33 condition: $V = 1.5, p < .001$; 83-17 condition: $V = 11, p < .001$).[6]

Finally, the change in within-pair difference between pre-interaction Recall 1 and post-interaction Recall 2 speaks to the lasting effects of interaction on participants' use of the singular marker. As can be seen from Figure 6, there is a small but statistically significant reduction in within-pair difference from Recall 1 to Recall 2 (n.s. difference between conditions,

[6] A reviewer asked if this reduction in within-pair difference reflects convergence within pairs, or if similar reductions in within-pair difference could arise as a by-product of most participants becoming independently consistent. To evaluate this hypothesis, we compared the mean within-pair difference in our data set at interaction block 2 (0.11) with the distribution of within-pair differences obtained by randomly shuffling participants across pairs. We generated 1000 pseudo-pairings by re-assigning participants to pseudo-pairs while respecting condition and participant (i.e. P1s from the 66-33 condition were only ever re-paired with P2s from the 66-33 condition) and measuring the mean within-pair difference at interaction block 2 in these new pseudo-pairings. The pseudo-pairings had reliably higher within-pair difference (the mean of the mean within-pair differences in 1000 randomisations was 0.46, and there were no cases where a random pseudo-pairing had mean within-pair difference equal to or lower than the mean of the veridical within-pair differences), indicating that this reduction in within-pair difference reflects genuine convergence in singular marking between interacting individuals.

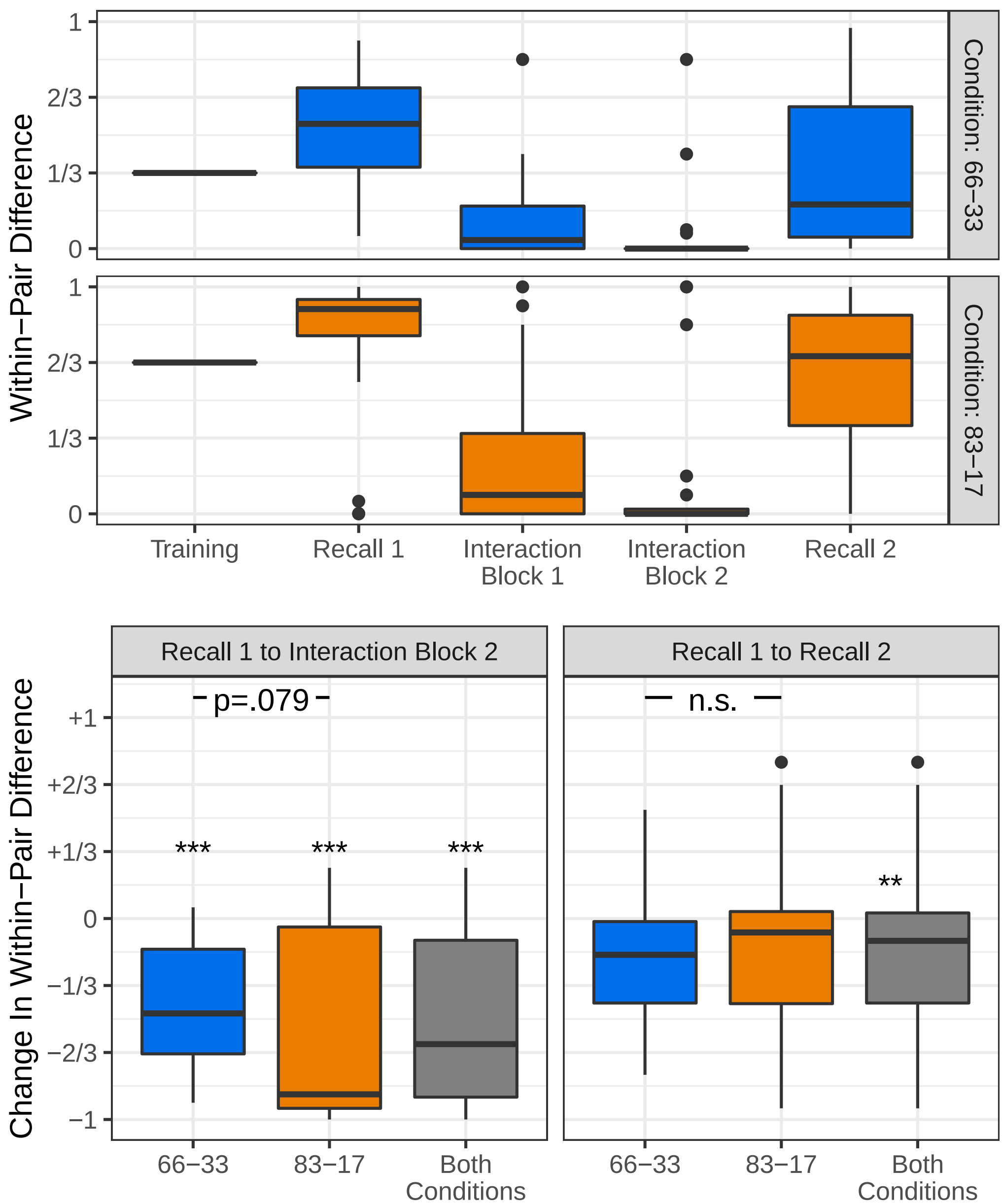

Figure 6: The upper panel shows the within-pair differences in marker use across the 5 stages of the experiment; the lower panels shows the change in those within-pair differences change from Recall 1 to interaction block 2, and from Recall 1 to Recall 2 — negative values for change indicate increased alignment between participants within a pair.

$W = 182.5, p = .646$; significantly non-zero change in within-pair difference, $V = 197, p = .007$), again providing some evidence that the effects of interaction persist beyond the duration of that interaction.

Discussion of Experiment 1

In Experiment 1 we trained participants on artificial languages exhibiting unpredictable variation in singular marking. In an individual recall test, participants on average produced the markers in a similar proportion as they occurred in their training language, although there was some evidence that participants were drawn somewhat towards extreme proportions. This finding is in line with previous research showing that adults are able to extract statistical properties from variable linguistic input (e.g. Hudson Kam & Newport, 2009; Perfors, 2016; Ferdinand et al., 2019), perhaps with some bias towards categoriality/regularity. Despite a tendency to produce variable marking, during the initial recall test, when participants used the language in a subsequent interaction task, they eliminated the variability and rapidly converged on systematic usage or non-usage of the marker. This is consistent with the results reported by Fehér et al. (2016), who show similar effects for artificial languages exhibiting unpredictably-variable word order. Previous research has shown that alignment does enhance communicative success (Pickering & Garrod, 2006), and that communicative design can affect local alignment (Branigan et al., 2011): the convergence to a common linguistic system in our study might therefore be because convergence better serves the purposes of interaction, in this case the correct identification of images.

Participants in Experiment 1 showed a preference for eliminating the singular marker, as evidenced by the overall drop in singular marking and the fact that P1s (trained on the higher frequency of marked singulars) showed greater reduction in singular marking than P2s. This could have been due to the fact that their native language, English, does not mark the singular. Alternatively, they might have noticed that it was more economical to omit the marker, since it was not necessary for disambiguation since plurals were always marked. In either case, this preference in Experiment 1 to eliminate singular marking provides an important contrast to the results of Experiment 2.

Finally, the post-interaction recall test provides some evidence that interaction had a small but lasting effect on participants' memory of their input language — these effects are quite variable, relatively small, and most pronounced in the individuals who change most during interaction (i.e. P1s, particularly in the 83-17 condition). In the general discussion we return to the question of whether a lasting effect of interaction is necessary for the regularising effects of interaction to play a direct role in language change.

Experiment 2

In Experiment 1, a change in marker use occurred very quickly during interaction, which could have been due to the fact that both participants in a pair were trained on a variable linguistic system, so when one of them dropped the marker, the other could follow suit without having to violate the rules of the grammar they had learnt during training. However, as discussed in the introduction, there are good reasons to expect that interaction will play out differently when one of the interacting individuals believes that marker use should be categorical, i.e. non-variable — if the grammar-based asymmetric accommodation hypothesis is correct, such individuals will be reluctant to change their behaviour to align with variable partners. Experiment 2 allows us test this hypothesis.

Method

Participants

Eighty-two participants were recruited from the University of Edinburgh's Student and Graduate Employment service and the University of Warwick's sign-up system for Psychology and Behavioural Science research to take part in an experiment that involves learning and interacting in a miniature artificial language. As in Experiment 1, participants were paid £8-10 for their participation (depending on time spent in the experimental booth).

Procedure

The procedure for Experiment 2 was identical to Experiment 1: participants were tested in pairs, worked through a computer program which presented and tested them on an artificial language, and then allowed them to use that language to communicate remotely with their partner.

Variable marking of the singular

As in Experiment 1, the training language provided post-nominal particles to mark singular and plural, with the plural consistently marked for all participants throughout training. As in Experiment 1 we manipulated the extent to which participants saw overt marking of the singular during training: participants either saw consistent categorical marking of the singular (100% marking), singular marking on 2 in 3 singulars (66% marking), or singular marking on 1 in 3 singulars (33% marking). For variably-trained participants, as in Experiment 1, the training data was constructed such that singular marking was unpredictable: every noun was marked for singular an equal number of times, and every verb appeared with a marked singular an equal number of times.

As in Experiment 1, participants within a pair differed in the language they were trained on. We ran 41 pairs: in the 100-66 condition (20 pairs), P1 was trained on 100% (categorical) marking, P2 was trained on 66% (variable) marking; in the 100-33 condition (21 pairs[7]), P1 was trained on categorical marking, P2 on 33% variable marking. These two conditions therefore both feature one categorically-trained participant and one variably-trained participant, with the difference in training frequency of marked singulars (33% difference in the 100-66 condition, 66% difference in the 100-33 condition), matched to the within-pair differences in Experiment 1.

Note that we make the categorical participants in every case categorical *users*, rather than non-users. This is a more conservative test of the grammar-based asymmetric accommodation hypothesis than using categorical non-users. Recall that Experiment 1 showed that participants tended to converge on non-marking of the singular, either because it is simply easier or due to interference from English (where the singular is unmarked). If we used categorical non-marking in Experiment 2 then any asymmetry in accommodation (which would in that case favour categorical non-marking) could be driven either by asymmetric accommodation *or* a preference to eliminate the redundant/non-English marker. In contrast, asymmetric accommodation to categorical *use* of the singular marker cannot be explained simply due to a more general tendency to omit the singular marker. Similarly, using categorical singular marking allows us to test whether the potential bias from English to drop the singular marker can be overcome in the right circumstances — again, any interference from English will tend to act against asymmetric accommodation in our experimental design, making this the more conservative test of our hypothesis.

[7]We aimed to collect 20 participant pairs per condition: however, due to the likelihood of no-shows by single participants, we over-booked participants, which generated one 'spare' pair. Rather than discard these extra participants, we include them in the analysis.

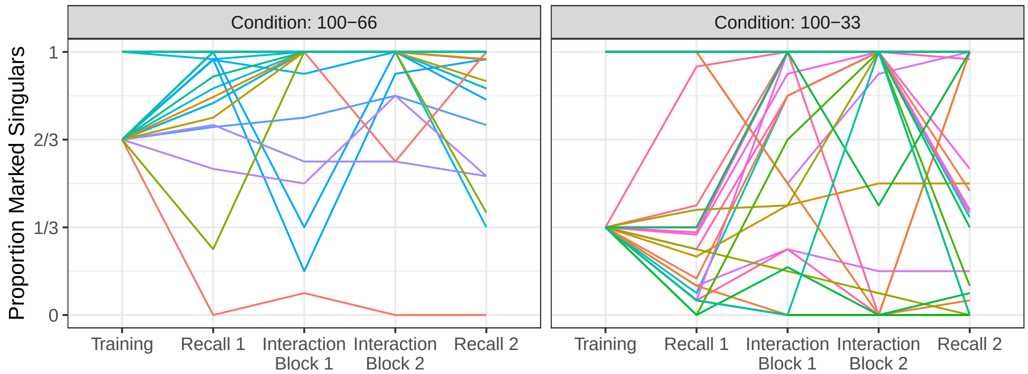

Figure 7: Proportion of trials in which the singular was marked, in training (determined by condition), Recall 1, interaction (split by block) and the post-interaction Recall 2. Each pair is represented by two lines, one per participant, sharing the same colour: alignment between participants is therefore reflected in lines of matching colour converging.

Analyses

The coding of participant descriptions was carried out through the same procedure as in Experiment 1, and our choice of non-parametric statistics on proportion data was motivated by the same concerns regarding convergence and non-normality.

Results

As in Experiment 1, performance during the communicative portion of the task was extremely high throughout, and varied little across conditions: the mean number of successful trials (in which the matcher selected the picture presented to the director) was 43.58 out of 48 in the 100-66 condition (42.9 in the first block of interaction, 44.25 in the second), 46.29 in the 100-33 condition (45.62 in block 1; 46.95 in block 2).

As in Experiment 1, our main dependent variable of interest is participants' use of the singular marker. Fig. 7 shows the full data for use of the singular marker across training, individual testing and two blocks of interaction (see also Fig. A.2 for separate by-pair plots). Fig. 8 provides means for the various phases.

In the 100-66 condition, categorically-trained participants remained categorical users of the singular marker throughout, barring two participants. One of them, interacting with a near categorical non-user, left the singular

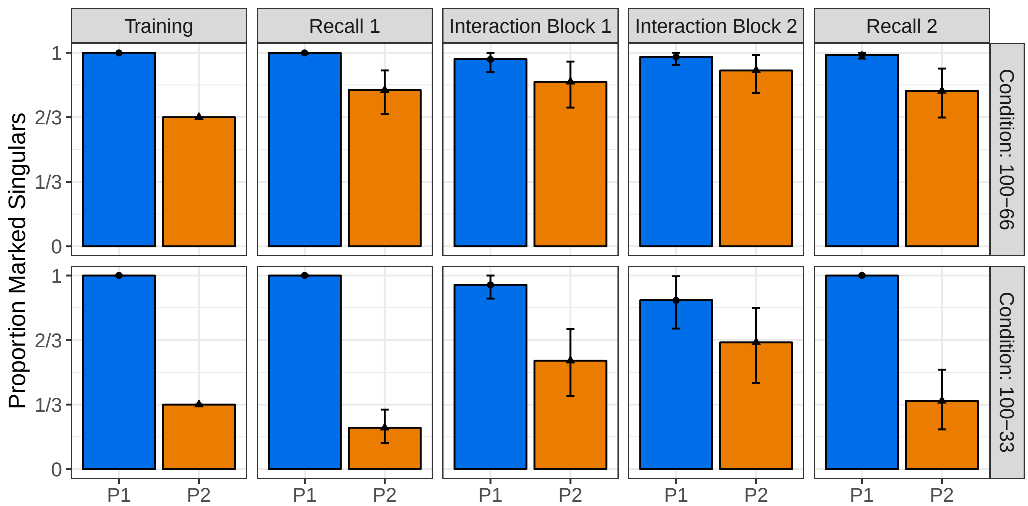

Figure 8: Mean proportion of trials in which the singular was marked in training (determined by condition), Recall 1, interaction (split by block) and the post-interaction Recall 2, for the 66-33 condition (upper panels) and 83-17 condition (lower panels). Error bars indicate bootstrapped 95% confidence intervals, obtained using 10,000 bootstrap samples and the percentile method.

unmarked on roughly half of the trials in interaction block 2. The other participant, in parallel with their partner's usage, dropped the marker in roughly 2/3 of the trials in interaction block 1 before becoming a categorical user again by block 2. Half of the variably-trained participants in this condition marked the singular variably during the pre-interaction recall test 1; during interaction, these variably-trained participants (with a few exceptions) rapidly aligned with their categorical partners, and remained largely categorical users in Recall 2.

In the 100-33 condition, we saw a similar pattern of results: the majority of categorically-trained participants remained categorical throughout (with only 4 of 21 becoming variable at some point during interaction, and all returning to categorical marking at Recall 2). Variable users in the 100-33 condition exhibited a spread of responses during individual testing, as was commonly the case in Experiment 1; during interaction, 13 of these participants accommodated upwards to become categorical users by the end of interaction.

In the following subsections we run through the same analyses as for Experiment 1, evaluating whether our participants probability matched during Recall 1, whether they changed their use of the singular marker during interaction and at Recall 2, relative to their use of the marker during Recall 1, and whether they aligned during interaction, i.e. came to use the singular marker in the same way as their partner. As in Experiment 1, we evaluate these questions using measures of change in frequency of use of the singular marker (see Figure 9) and within-pair difference (see Figure 10). We then present additional analyses speaking to the grammar-based asymmetric accommodation hypothesis.

Change in marker usage

Figure 9 plots the change in marker usage across three key phases of Experiment 2, comparing proportion of marked singulars produced during Recall 1 to that seen during training (upper panel); change from Recall 1 to interaction block 2 (middle panel), and change from Recall 1 to Recall 2 (lower panel).

As in Experiment 1, the change in frequency of singular marking between participants' training data and their productions in Recall 1 indexes the extent to which participants reproduced the frequency of singular marking seen in their training data, with values around 0 indicative of probability matching. Categorically-trained participants were clearly highly accurate in

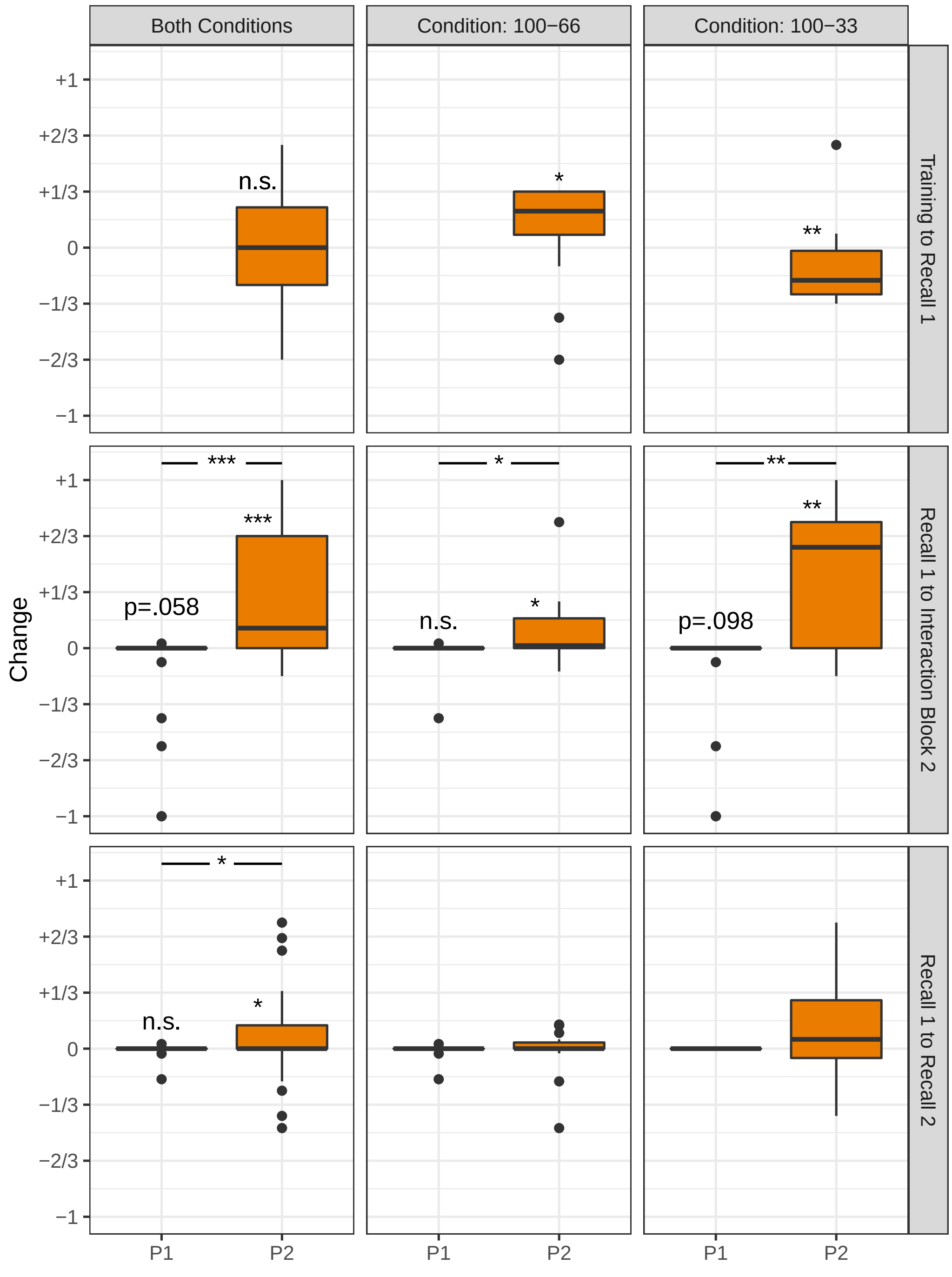

Figure 9: Experiment 2, change in proportion of marked singulars from training to Recall 1 (upper figure), from Recall 1 to block 2 of Interaction (middle figure) and from Recall 1 to Recall 2 (right figure). In all cases, change is calculated as proportion of marked singulars at the later stage of the experiment minus proportion of marked singulars at the earlier stage - i.e. positive values indicate an increase in singular marking, negative values indicate a decrease.

reproducing the singular marking seen in their training data — all but one participant marked the singular categorically during Recall 1 (that participant omitted the singular marker once), and for this reason were excluded from this analysis. Among the variably-trained participants, there is a difference between conditions ($W = 347.5, p < .001$): while the complete dataset suggests probability matching (change is not different from 0 when collapsing across conditions, $V = 425, p = .630$), variably-trained participants in the 100-66 condition on average produced marked singulars slightly above that of their input data ($V = 167, p = .021$), while variably-trained participants in the 100-33 condition under-produced the marked singular, as shown by a non-zero difference between training and Recall 1 ($V = 28, p = .007$). This mirrors the pattern we see in Experiment 1, where variably-trained participants are pulled slightly towards the extremes of singular marking, although in Experiment 1 this effect was clearest in participants trained on more extreme proportions (i.e. in the 83-17 condition).

The change in frequency of singular marking between Recall 1 and interaction block 2 (middle panel of Figure 9) shows a pattern of results which are strikingly different to those seen in Experiment 1, and consistent with the asymmetric accommodation hypothesis. Recall that in Experiment 1 we saw an overall reduction in singular marking, driven by the tendency of P1 participants (trained variably, but on more frequent use of the singular marker than their partner) to reduce their use of the singular. In contrast, here we see the reverse pattern, where participants trained on the less frequent, variable use increase their usage of the singular marker during interaction; as can be seen in Fig. 7, unlike in Experiment 1, most pairs converged during interaction on systems in which the singular was always marked. Collapsing across conditions (the effect of condition is n.s., $W = 804, p = .720$), P1 and P2 differ in the amount of change they show between Recall 1 and interaction block 2 (as indicated by a significant effect of the P1/P2 contrast, $W = 421.5, p < .001$); there is also an interaction between condition and P1/P2 ($W = 341, p < .001$), suggesting the difference in the behaviour of P1 and P2 differs between conditions. Taking our data set as a whole and collapsing across condition: whereas the categorically-trained participants did not reliably change their usage of the singular marker during interaction (mean change is only marginally different from zero: $V = 1, p = .058$, driven by 5 out of 41 categorically-trained participants who reduced their marker use during interaction), variably-trained participants reliably *increased* their usage of the singular marker ($V = 477.5, p < .001$). This same pattern of results is borne out in an analysis considering each condition separately, motivated by the condition by P1/P2 interaction: both conditions show significant differences between P1s and P2s in amount of change from Recall 1 to interaction block 2 (100-66 condition: $W = 118, p = .010$; 100-33 condition: $W = 95.5, p = .001$), and in both conditions P2s show a significant increase in singular marking whereas P1s do not show a reliable effect (100-66 condition, P2s: $V = 70.5, p = .015$; 100-33 condition, P2s: $V = 189.5, p = .002$; 100-66 condition, P1s: $V = 1, p = 1$; 100-33 condition, P1s: $V = 0, p = .098$). The interaction between condition and P1/P2 is driven by the fact that the change by P2s is clearly larger in the 100-33 condition than in the 100-66 condition, as they have further to move to accommodate to their categorical partners (P1s do not differ in amount of change according to condition, $W = 249, p = .102$; P2s do differ across conditions in amount of change, $W = 133, p = .045$).

Finally, analysis of the change in singular marking from Recall 1 (pre-interaction) to Recall 2 (post-interaction) suggests a similar picture to that seen in Experiment 1. In Experiment 1 there was some evidence of a lasting effect of interaction: the participants who were trained on the more frequent use of the singular marker changed (reduced) their use of the singular marker more during interaction, and then persisted in under-producing (relative to Recall 1) at Recall 2 (where they were asked to recall the initial language they were trained on). In Experiment 2 we see a similar pattern, in that the participants who changed most during interaction (here the P2s) showed some evidence of lasting effects. The difference between conditions in our Experiment 2 data was not significant ($W = 757, p = .388$), there was a significant difference between P1 and P2 ($W = 624, p = .024$) and no interaction between condition and P1/P2 ($W = 250.5, p = .293$), suggesting this P1/P2 difference was roughly equivalent across conditions. While the overall dataset (i.e. including both P1s and P2s; note that P1s were predicted to not change their marker use during interaction and therefore not predicted to show post-interaction effects) showed only a marginal change from Recall 1 to Recall 2 ($V = 386.5, p = .059$), P1s and P2s showed different patterns of behaviour: P1s showed no lasting change from Recall 1 to Recall 2 ($V = 1, p = .423$), whereas P2s significantly increased their marker usage ($V = 338, p = .031$). This supports our hypothesis, which predicts lasting accommodation in variable users towards categorical users but not vice versa.

Change in within-pair differences

As in Experiment 1, and as can be seen in Figure 7, there is a strong tendency for pairs of participants to converge on a shared system of using the singular marker. Figure 10 plots within-pair difference across the various stages of the experiment, as well as the change in within-pair difference at several key stages.

Within-pair differences sharply reduced during interaction, as is clear from the lower left panel of Figure 10 showing change in within-pair difference from Recall 1 to interaction block 2. While this effect is evident across the entire data set, collapsing across conditions ($V = 31, p < .001$), it is most pronounced in the 100-33 pairs and there is a significant difference between conditions in the amount of change in within-pair difference from Recall 1 to interaction block 2 ($W = 340.5, p < .001$). However, both conditions independently show a significant reduction in within-pair difference (100-66 condition: $V = 7.5, p = .009$; 100-33 condition: $V = 8.5, p < .001$).

Finally, the lower right panel of Figure 10 provides additional (weak) evidence to support the earlier analysis suggesting lasting effects of interaction, here meaning that participants are more similar in their post-interaction marker use than in their pre-interaction use: there is a marginal reduction in within-pair difference from Recall 1 to Recall 2 (n.s. difference between conditions, $W = 264.5, p = .156$; collapsing across conditions, change in within-pair difference is marginally different from 0, $V = 158, p = .079$).

Grammar-based asymmetric accommodation

As discussed above, the results of Experiment 2 are as predicted by the grammar-based asymmetric accommodation hypothesis — despite the general preference seen in Experiment 1 for the participant trained on more frequent singular marking to reduce the frequency of singular marking during interaction, categorically-trained participants in Experiment 2 do not reliably do so (their change in proportion of marked singulars from Recall 1 to interaction block 2 only differs marginally from 0, $V = 1, p = .058$), forcing their variably-trained partners to align upwards by increasing their use of the singular marker (their partners' change from Recall 1 to interaction block 2 is positive and significantly different from 0, $V = 477.5, p < .001$), and consequently categorically-trained participants differed significantly from their variably-trained partners in the extent to which they changed their behaviour during interaction (collapsing across condition, P1s and P2s differ significantly in the extent to which they change their marker use from recall 1

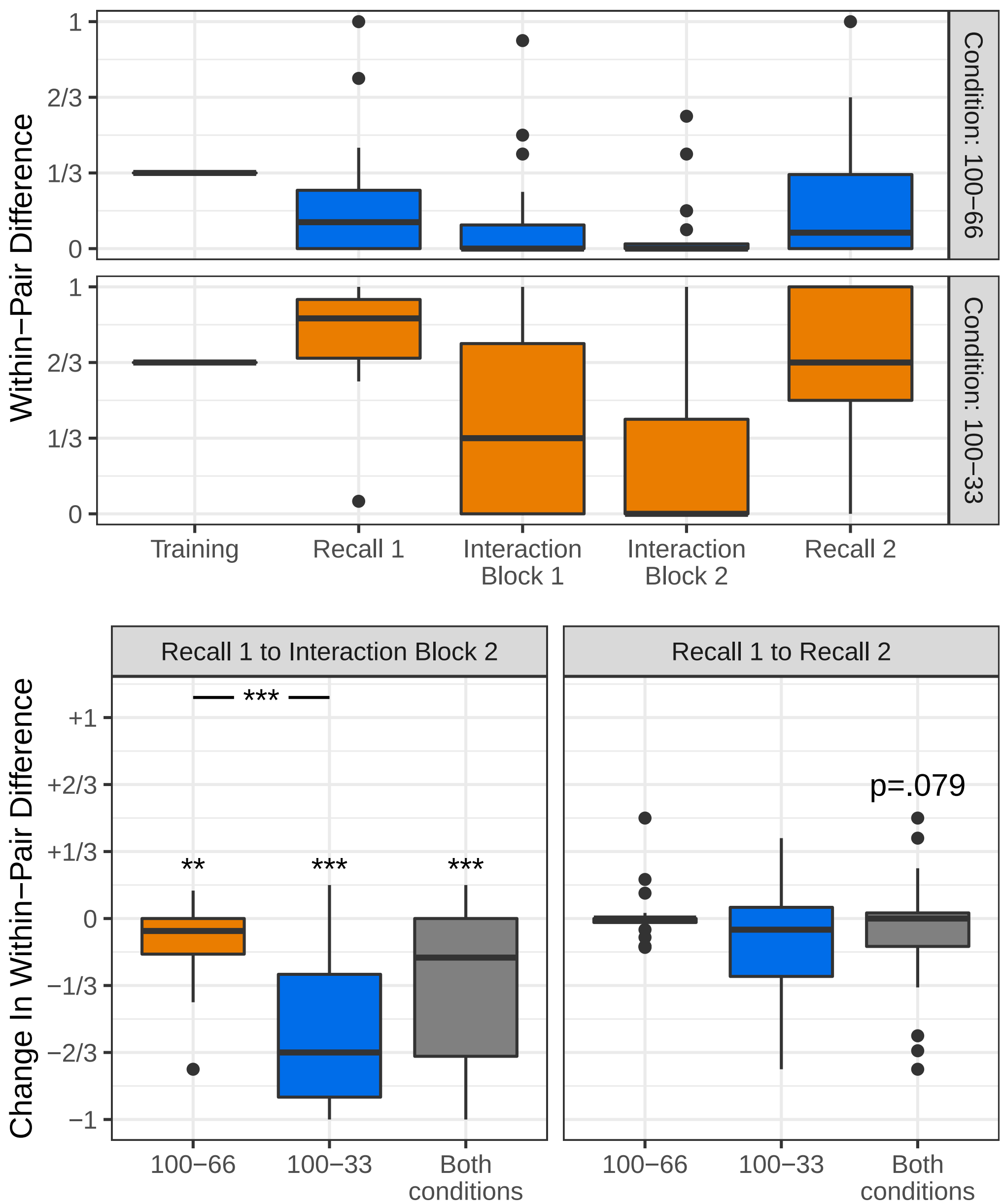

Figure 10: The upper panel shows the within-pair differences in marker use across the 5 stages of Experiment 2; the lower panels show the change in those within-pair differences change from Recall 1 to interaction block 2, and from Recall 1 to Recall 2 — negative values for change indicate increased alignment between participants within a pair.

to interaction block 2, $W = 421.5, p < .001$). Further evidence of this effect can be obtained by combining the data from variably-trained participants across Experiment 1 and Experiment 2 (see Figure 11, upper left panel) — these groups significantly differ in their change in use of the marked singular between Recall 1 and interaction block 2 ($W = 2478.5, p < .001$), with Experiment 1 participants (paired with a variably-trained partner) decreasing their use and Experiment 2 participants (paired with a categorically-trained partner) increasing their use. This same pattern of results holds if we look only at P2 participants, i.e. those participants who were paired with a partner who was trained on more frequent use of the singular marker: P2s differ in the amount of change from Recall 1 to interaction block 2, depending on whether they were paired with a variable or categorical partner, $W = 1123, p = .004$ (see Figure 11, upper right panel). As a last comparison, we can compare participants who were trained on 33% marked singulars in the 66-33 condition with those trained on the same proportion of marked singulars but paired with a categorical partner (in the 100-33 condition): these two groups of participants, who received identical training and were paired with a partner who used the plural more frequently than themselves, differ significantly in their change in the use of the singular between Recall 1 and interaction block 2, depending on whether their partner was trained on categorical or variable singular marking ($W = 311, p = .009$; see Figure 11, lower left panel).

Finally, we can ask whether this difference in behaviour of categorically-trained participants is due to their categorical training, or their categorical production of the singular marker during Recall 1 (which presumably reflects their belief that singular marking should be categorical). The change from Recall 1 to interaction block 2 for all participants who produced 100% marked singulars at Recall 1 is shown in the lower right panel of Figure 11, split according to whether their training was variable (N=22 out of 120 variably-trained participants) or categorical (N=40). The mode change for both groups is 0: while the variably-trained participants seem slightly more likely to radically change their behaviour during interaction (5 of 22 variably-trained participants became categorical non-users during interaction, 5 of 41 categorically-trained participants became non-categorical, but only 2 of those became categorical non-users), this difference is not statistically significant ($W = 492.5, p = .231$). This suggests that the participant's *belief* that the singular marker should be used categorically is the main driver of the asymmetric accommodation effect, rather than absence of variation in their training input.

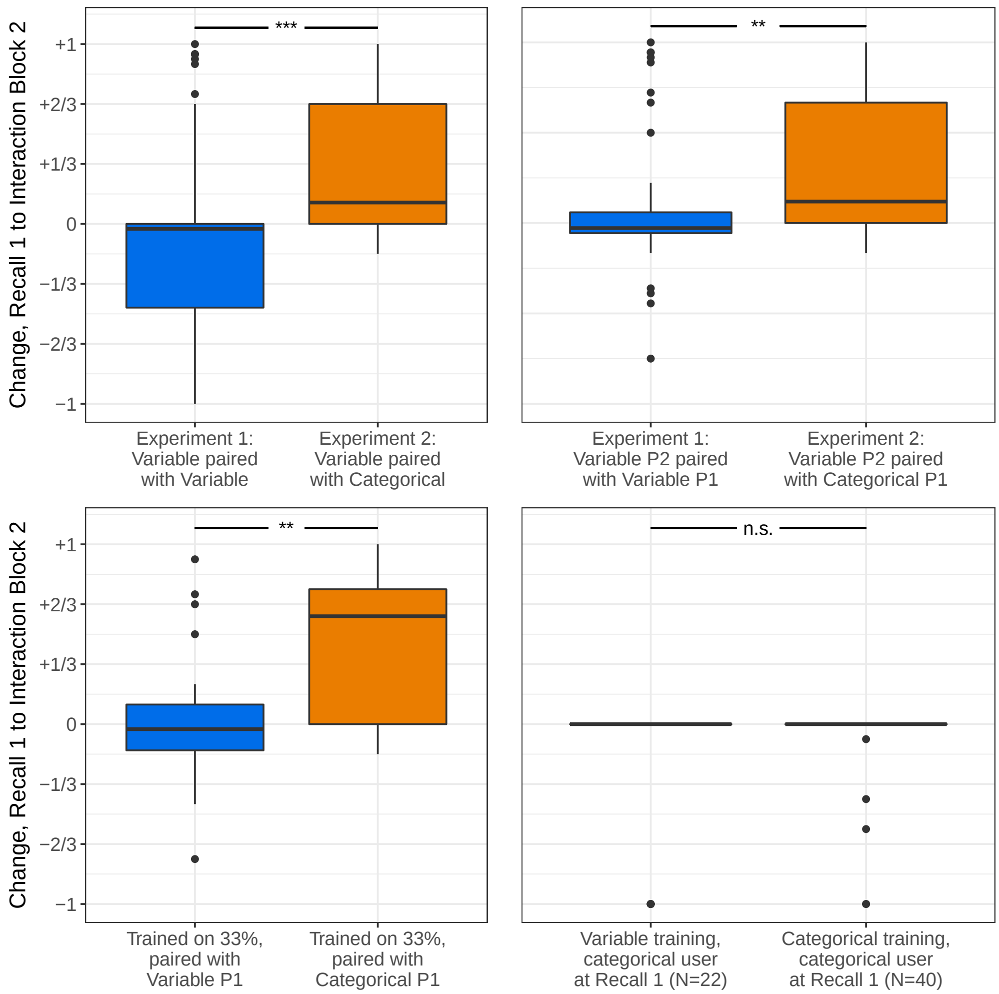

Figure 11: Change from Recall 1 to interaction block 2 for all variably-trained participants across the two experiments (upper left panel), for all P2s (who are trained variably and on a lower frequency of marked singulars than their partners, upper right panel) and for participants who were trained on 33% marked singulars (lower left panel). In all cases, participants who were paired with categorical partners behaved differently from participants who were paired with variably-trained partners. The lower right panel shows change from Recall 1 to interaction block 2 for participants who produced 100% (categorical) singular marking in Recall 1, split according to whether their training was variable or categorical.

General Discussion

We presented two experiments investigating the effects of communicative interaction on unpredictably variable linguistic systems. We found that unpredictable variation was greatly reduced or eliminated during interaction, and the effects of interaction persisted into a post-interaction recall test (in both experiments, a point to which we return below). Importantly, our data are consistent with the grammar-based asymmetric accommodation hypothesis, which states that variable users are more likely to adapt their linguistic behaviour to categorical users rather than vice versa. These results speak to a number of larger issues regarding diachronic linguistic change and language evolution.

Additional thoughts on the grammar-based asymmetric accommodation hypothesis

As predicted by the grammar-based asymmetric accommodation hypothesis, we found an asymmetry in the behaviour of participants trained on variable vs. categorical linguistic systems. Categorically-trained participants used singular markers according to the rule that all singulars had to be marked. Even though they were exposed to unmarked singulars when they interacted with their variable partners, for the most part they did not accommodate to them but maintained their deterministic usage. Variably-trained participants, on the other hand, were much more likely to adopt the system of their categorical partners, even though — as shown in Experiment 1 — marking the singular is against participants' natural tendency to drop the marker when that option is available, either due to native language influences, minimisation of effort (as show in other artificial language learning/interaction experiments, e.g. Fedzechkina et al., 2016; Kanwal et al., 2017), or other biases in learning or perception. Despite quickly adopting categorical usage during interaction, the participants who inferred a variable grammar remained aware that the system allows for variability, as confirmed by their variable output during post-interaction recall tests.

The grammar-based asymmetric accommodation hypothesis explains this asymmetry in terms of the difference in underlying grammars for variably- and categorically-trained participants: since variable users did not have to violate the rules of the grammar they had inferred during training, they were more likely to accommodate to their categorical partners. This suggests that at least three pressures are at play in shaping alignment between interlocutors in our experiments: a preference to align with one's interlocutors (evident in the behaviour of virtually all variably-trained participants), a preference to minimise production effort (evident in Experiment 1 in the tendency to drop the redundant singular marker), and a preference to use forms that are permitted under the inferred grammar, even if those forms are assigned low probability (leading to the asymmetric accommodation effects seen in Experiment 2).

Our experimental data suggest an additional factor that may contribute to this asymmetry between variable and categorical use in interaction. We found that convergence by variably-trained participants to their categorically-trained partners happened rapidly. Therefore, categorical users had little opportunity to even notice the absence of singular markers in the communicative behaviour of their partners, and if they did, they might have dismissed initial omissions as isolated errors. Either way, this would have decreased the probability that categorical users should be influenced by unmarked singulars in the output of their partners. The rapidity of convergence might therefore contribute to the explanation for why accommodation favours categorical usage over pragmatically-conditioned usage: rapid convergence means that there simply is not enough time to realise that one's partner uses a given form variably, let alone to infer the pragmatic subtleties conditioning its use. To a categorical user, variability might appear unsystematic at first even if it in fact depends on pragmatic conditions in a predictable fashion, such as how much one speaker thinks the other one knows already, how much inferencing work polite speakers can expect of their addressees, and how polite they want to be in the first place. While is it clearly possible to identify such conditioning factors (after all, language learners do eventually acquire even complex rules of variable pragmatic conditioning), it may require a lot of evidence, making it hard to achieve in a couple of minutes during a single interaction. Thus, quick attempts by categorical users to emulate the variable usage of their interlocutors are likely to fail, while the reverse does not hold: it should be relatively easy to figure out that a speaker uses a constituent whenever the grammar allows it. Therefore, usage patterns that are grammatically and categorically conditioned can be emulated quickly. Once they are emulated, however — i.e. as soon as variable users begin to accommodate to categorical interlocutors — the latter will be deprived of evidence for the conditions behind variable use.

This discussion of the challenges imposed by acquiring conditioned variation during interaction also highlights a mismatch between our experimental design and the cases of obligatorification that inspired it: namely, in the Old English case we discuss, use of the demonstrative was pragmatically-conditioned, rather than (as in our variable training languages) unconditioned. This seems to us a reasonable first step in demonstrating asymmetric accommodation, and in other work we find the same asymmetric accommodation effects when one member of a pair learns a system of lexically-conditioned (rather than unconditioned) variation (Atkinson et al., 2018). This provides at least one demonstration that asymmetric accommodation can lead to convergence on categorical systems at the expense even of conditioned systems of variation; it would of course be worthwhile to test whether there are any limits to this (e.g. if highly entrenched systems of conditioned variation will similarly be abandoned in interaction), and whether the transparency of the conditioning factors to the naive categorical participant affects the alignment process (in particular, whether more 'obvious' conditioning patterns are more likely to survive interaction). In this connection, it would be satisfying to also look at the case of pragmatically-conditioned variation, which we expect to be relatively non-transparent and therefore prone to elimination during interaction.

Finally, we unexpectedly found in Experiment 2 that variably-trained participants who behaved as categorical users in the pre-interaction recall test also seemed to stick to their deterministic usage of the singular marker during interaction. While this conclusion rests on a null finding in an unbalanced dataset using relatively weak non-parametric tests, and should therefore be treated with caution, this suggests that once a linguistic rule is internalised, people are reluctant to deviate from it, unless they interpret variability to be part of the linguistic rule. In other words, it is the grammar that the learner infers, that determines the asymmetry, rather than the input the grammar was inferred from.

Does interaction have lasting effects on variability?

Our experimental data provides some evidence that the reduction in variability seen during interaction persists beyond that interaction, specifically into the post-interaction recall test. In both experiments, the participants who accommodates most to their partners during the interaction phase (P1s in Experiment 1 and the variably trained P2s in Experiment 2) exhibited a lasting change in their use of the singular marker; these effects were most visible in the individuals who changed the most during interaction (P1s in the 83-17 condition in Experiment 1, who tended to substantially reduce their marker use to align with their less-frequently-marking partner; P2s in the 100-33 condition in Experiment 2, who had to substantially increase their marker use to accommodate to their categorical partner), although our statistical tests indicated that the lasting effects of interaction were roughly equivalent across conditions. We see similar lasting effects of interaction in other artificial language learning paradigms (across two experiments in Fehér et al., 2016). However, in the experiments reported here these effects are generally small and quite variable across participants, which warrants further discussion.

Firstly, at the start of Recall 2 participants were instructed to recall the original language they were trained on. This means that our method for measuring the lasting effect of interaction is (intentionally) quite conservative: we were looking for effects sufficiently strong to survive an explicit instruction to revert to an earlier behaviour. Alternative approaches to this post-interaction recall test may yield clearer evidence of lasting effects. For instance, more neutral instructions prior to an asocial recall test, or a second phase of communicative interaction with a new partner, would allow participants more freedom to persist in the behaviour they adopted during interaction. Given that we see some evidence of persistent effects even given our very conservative framing, we expect that such effects would be more apparent using those methods. It is also likely that any lasting effects on individual linguistic behaviour will depend on other factors, such as relative social status and the number of interlocutors one has interacted with, factors which we don't manipulate here.

Secondly, there is some question of whether lasting effects of interaction are actually required for changes operating during interaction to propagate through a population. Lasting effects on individual behaviour may not be required to drive language change: for instance, children learn their language by participating in and observing interactions, including interactions between other adults and older peers, and they might well be influenced by modifications which only last for the duration of a specific interaction. If interlocutors become less variable for the course of an interaction, they would suppress evidence for variability for any child acquiring their language through observing or participating in that particular interaction. This means that modifications occurring during interaction could have lasting influences on the population's behaviour even if those modifications are themselves fleeting. However, the propagation of linguistic changes is likely to be more rapid if the effects of interaction on an individual's behaviour outlast the duration of that interaction, and larger lasting effects should lead to faster changes. It may be that small post-interaction effects such as we see in our experiment will simply be swamped by other factors when individuals are embedded in populations.

Mechanisms of regularisation

Previous research identified two ways in which regular linguistic systems may emerge from unpredictably-variable starting points. Regularity may be a product of relatively strong biases in learning in individuals (e.g. Hudson Kam & Newport, 2009; Perfors, 2016), or may emerge more gradually through transmission (e.g. Smith & Wonnacott, 2010; Reali & Griffiths, 2009). Our experiments identify an additional mechanism: communicative interaction. We find that interaction leads to a reduction in variation, as also shown by Fehér et al. (2016). Grammar-based asymmetric accommodation further helps to explain the establishment of categorical usage patterns in speech communities. Since languages are conventional, socially shared systems, one cannot fully explain their properties by asking how easily they are acquired by individuals; one also needs to ask how easily they are shared. Our experiments have revealed asymmetries that bias the direction of accommodation in interaction, and that may help to explain why in the historical record categorical usage patterns tend to oust variable ones once they emerge in a population of speakers. More generally, these biases may also help to explain why the grammaticalisation processes attested in the histories of practically all natural languages appear to be unidirectional and irreversible.

Our results do not imply that variable usage patterns will generally be ousted by competing categorical ones. As far as grammaticalisation is concerned, it is known to be cyclic, and constituents that become obligatory in one phase may become optional again in later phases. Articles are themselves a case in point: deriving from optionally used numerals or demonstratives, they come to be obligatory in specific syntactic contexts. In later phases, however, they may grammaticalise further into general noun phrase markers (Himmelmann, 2001; Greenberg, 1978), which are semantically empty but highly frequent. Therefore, they become once more prone to phonological reduction and deletion, i. e. they become optional (again) before possibly being lost altogether. The dynamics driven by the asymmetric accommodation bias revealed in Experiment 2 are obviously characteristic only of those specific phases in grammaticalisation in which variable use becomes categorical; our experiments help to explain why accommodation may indeed lead to the elimination of variation under such circumstances.

Future directions

In addition to the questions raised above, a number of other questions remain to be addressed. Firstly, we have only considered presence/absence variation: other paradigms (e.g. Hudson Kam & Newport, 2009; Smith & Wonnacott, 2010) look at variation where there are two or more overt markers for a single function, and it may be that alignment during interaction proceeds differently in such cases. Secondly, we look only at alignment within pairs who undergo a relatively short period of training and a relatively long, intense period of interaction with a single partner: since the real-world case involves longer learning (perhaps entailing greater commitment to the trained system) and interaction with a wider range of partners, this seems like a worthwhile scenario to explore experimentally.

Finally, accommodation is surprisingly rapid in our study: a great deal of alignment takes place in the first few trials of interaction. It would be intriguing to investigate the lower-level processes by which participants come to decide how to use markers after just one or two exposures to the marking behaviour of their partner. Similarly, one might ask how that might change if one increases the knowledge that participants have of the language used by their partner. It might make a difference, for example, if participants are trained together rather than — as in our experiments — in isolation.

Conclusions

Accommodation during interaction leads to the elimination of unpredictable variation and consequently provides an additional (complementary) mechanism for explaining the absence of unpredictable variation in natural languages. In line with historical evidence, accommodation seems to be inherently asymmetric. While variable users accommodate to categorical partners by increasing their frequency of usage, categorical users do not tend to accommodate to variable partners by becoming variable. Thus, when, in a population, the number of speakers who use a marker categorically reaches a critical threshold, asymmetric accommodation may drive the population towards uniformly categorical marker use. The grammar-based asymmetric accommodation hypothesis therefore offers a potential mechanistic explanation for the recurring tendency for obligatorification during language change, which is central to attested changes such as the emergence of the definite article in English, as well as to processes of grammaticalisation more generally.

*

Acknowledgments This work was supported by a Newton International Fellowship, held by O. F., and by the Economic and Social Research Council (grant number ES/K006339, held by K. S). This project has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 681942, held by K. S.

References

Aitchison, J. (1996). Small steps or large leaps? undergeneralization and overgeneralization in creole acquisition. In H. Wekker (Ed.), *Creole languages and language acquisition* (pp. 9–31). Berlin, Germany: Mouton de Gruyter.

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of conditional probability statistics by human infants. *Psychological Science*, *9*, 321–324.

Atkinson, M., Smith, K., & Kirby, S. (2018). Adult learning and language simplification. *Cognitive Science*, *42*, 2818–2854.

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

Bayley, R. (1996). Competing constraints on variation in the speech of adult chinese learners of english. In R. Bayley, & D. R. Preston (Eds.), *Second Language Acquisition and Linguistic Variation* (pp. 97–120). John Benjamins.

Bickerton, D. (1981). *Roots of Language*. Ann Arbor, MI: Karoma.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences*, *7*, 173–221.

Bickerton, D., & Givón, T. (1976). Pidginization and syntactic change: From sxv to vsx to svx. In S. B. Steever, C. A. Walker, & S. Mufwene (Eds.), *Papers from the parasession on diachronix syntax* (pp. 9–39). Chicago Linguistic Society.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: evidence from dialogs with humans and computers. *Cognition*, *121*, 41–57.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 1482–1493.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*, 711–733.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge University Press.

Bybee, J., & Beckner, C. (2009). Usage based theory. In B. Heine, & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 828–855). Oxford: Oxford University Press.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2010). A vowel is a vowel: generalizing newly learned phonotactic constraints to new contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 821–828.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317–335.

Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language*, *46*, 49–70.

Coupland, N. (2010). Accomodation theory. In J. Jaspers, J. Verschueren, & J. Östman (Eds.), *Society and language use* (pp. 21–27). Amsterdam: Benjamins.

Croft, W. (2000). *Explaining Language Change: an evolutionary approach*. London: Longman.

DeGraff, M. (Ed.) (1999). *Language Creation and Language Change: Creolization, diachrony, and development*. Cambridge, MA: MIT Press.

Diewald, G., & Smirnova, E. (2010). Paradigmaticity and obligatoriness of grammatical categories. *Acta Linguistica Hafniensia*, *42*, 1–10.

Dixon, R. M. W. (1972). *The Dyirbal language of North Queensland*. Cambridge: Cambridge University Press.

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. C. (2016). The relationship between artificial and second language learning. *Cognitive Science*, *40*, 822–847.

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2016). Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science*, *41*, 416–446.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180.

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68.

Fischer, J. L. (1958). Social influences on the choice of a linguistic variant. *Word*, *14*, 47–56.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*, 218–245.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.

Garrod, S., & Pickering, M. J. (2013). Dialogue: Interactive alignment and its implications for language learning and language change. In P. Binder, & K. Smith (Eds.), *The Language Phenomenon: Human communication from milliseconds to millennia* (pp. 47–64). Berlin: Springer.

Garrod, S. C., & Clark, A. (1993). The development of dialogue co-ordination skills in schoolchildren. *Language and Cognitive Processes*, *8*, 101–126.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249–268.

Giles, H. (Ed.) (1984). *The dynamics of speech accommodation* volume 46. International Journal of the Sociology of Language.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, & C. N. (Eds.), *Contexts of Accommodation* (pp. 1–68). Cambridge University Press.

Giles, H., & Ogay, T. (2007). Communication accommodation theory. In B. B. Whaley, & W. Samter (Eds.), *Explaining Communication: Contemporary Theories and Exemplars*. Mahwah, NJ: Lawrence Erlbaum.

Givón, T. (1985). Function, structure, and language acquisition. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition* (pp. 1005–1028). Hillsdale, NJ: Lawrence Erlbaum volume 2.

Greenberg, J. H. (1978). How does a language acquire gender markers? In J. H. Greenberg (Ed.), *Universals of human language, Vol. 3: Word structure* (pp. 47–82). Stanford: Stanford University Press.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, *34*, 365–399.

Heine, B. (1997). *Cognitive foundations of grammar*. Oxford: Oxford University Press.

Himmelmann, N. (1997). *Deiktikon, Artikel, Nominalphrase: Zur Emergenz syntaktischer Struktur*. Tübingen: Niemeyer.

Himmelmann, N. (2001). Articles. In M. Haspelmath, E. Koenig, & W. Oesterreicher (Eds.), *Language typology and language universals* (pp. 831–841). Berlin, Germany: Mouton de Gruyter.

Hockett, C. F. (1963). The problem of universals in language. In J. H. Greenberg (Ed.), *Universals in language* (pp. 1–29). Cambridge, MA: MIT Press.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. In D. Schiffrin (Ed.), *Meaning, Form, and Use in Context: Linguistic Applications* (pp. 11–42). Washington, DC: Georgetown University Press.

Hudson Kam, C. L. (2009). More than words: Adults learn probabilities over categories and relationships between them. *Language Learning and Development*, *5*, 115–145.

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

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

Ibbotson, P. (2013). The scope of usage-based theory. *Frontiers in psychology*, *4*.

Johnson, J. S., Shenkman, K. D., Newport, E. L., & Medin, D. L. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language*, *35*, 335–352.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

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

Labov, W. (1963). The social motivation of a sound change. *Word*, *19*, 273–309.

Lass, R. (1984). *Phonology: An introduction to basic concepts*. Cambridge: Cambridge University Press.

Leafgren, J. (2002). *Degrees of explicitness: Information structure and the packaging of Bulgarian subjects and objects*. Amsterdam: John Benjamins.

Leech, G. (2014). *The pragmatics of politeness*. New York, NY: Oxford University Press.

Lehmann, C. (1985). Grammaticalization: Synchronic variation and diachronic change. *Lingua e Stile*, *20*, 303–318.

Lieven, E. (2014). First language development: a usage-based perspective on past and current research. *Journal of Child Language*, *41*, 48–63.

Lieven, E., & Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective. In N. Ellis (Ed.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 168–196). New York: Routledge.

Loy, J., & Smith, K. (submitted). Syntactic adaptation may depend on perceived linguistic knowledge: Experimental evidence from interaction with native and nonnative speakers.

MacLeod, B. (2012). *The effect of perceptual salience on phonetic accommodation in cross-dialectal conversation in Spanish*. Ph.D. thesis University of Toronto Toronto.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word recognition and learning: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, *132*, 202–227.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development*, *12*, 138–155.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–225.

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*, 203–228.

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

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–328.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*, 30–54.

Reinöhl, U. (2016). *Grammaticalization and the rise of configurationality in Indo-Aryan*. Oxford: Oxford University Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Shuy, R. W., Wolfram, W., & Riley, W. K. (1967). *Linguistic correlates of social stratification in Detroit speech*. Technical Report U. S. Office of Education Final Report No. 6-1347.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, *49*, 370–407.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use, and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B*, *372*, 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*, 444–449.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Soliz, J., & Giles, H. (2014). Relational and identity processes in communication: a contextual and meta-analytical review of communication accommodation theory. In E. L. Cohen (Ed.), *Communication yearbook 38* (pp. 107–145). Routledge.

Sommerer, L. (2011). *Old English se: From demonstrative to article. A usage-based study of nominal determination and category emergence*. Ph.D. thesis University of Vienna Vienna, Austria.

Sommerer, L. (2012). Investigating the emergence of the definite article in Old English: About categorization, gradualness and constructions. *Folia Linguistica Historica*, *33*, 175–213.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.*. Cambridge, MA: Harvard University Press.

Trudgill, P. (2008). Colonial dialect contact in the history of european languages: On the irrelevance of identity in new-dialect formation. *Language in Society*, *37*, 241–254.

van de Velde, F. (2010). The emergence of the determiner in the dutch np. *Linguistics*, *48*, 263–299.

Vihman, V. A., Nelson, D., & Kirby, S. (2018). Animacy distinctions arise from iterated learning. *Open Linguistics*, (pp. 552–565).

Vincent, N. (1997). The emergence of the d-system in romance. In A. van Kemenade, & N. Vincent (Eds.), *Parameters of morphosyntactic change* (pp. 179–274). Cambridge: Cambridge University Press.

Wedel, A. (2007). Feedback and regularity in the lexicon. *Phonology*, *24*, 147–185.

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language and Communication*, *2*, 57–89.

Wolfram, W. (1985). Variability in tense marking: A case for the obvious. *Language Learning*, *35*, 229–253.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*, 1–14.

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, *66*, 458–478.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*, 165–209.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

*

Appendix: By-pair plots of singular marking

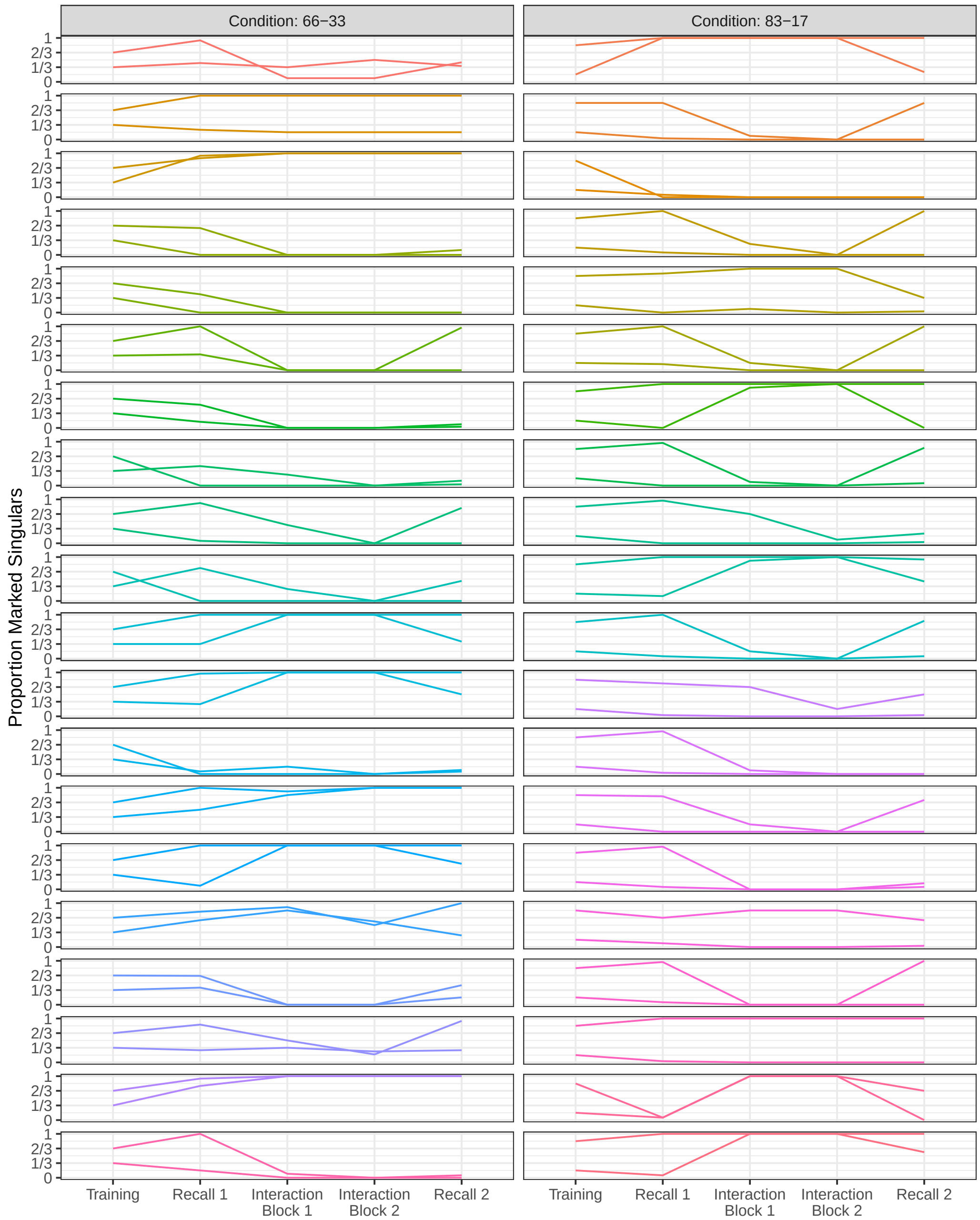

Figure A.1: Experiment 1. Proportion of trials in which the singular was marked, in training (determined by condition), Recall 1, interaction (split by block) and the post-interaction Recall 2, with separate plots showing each pair individually.

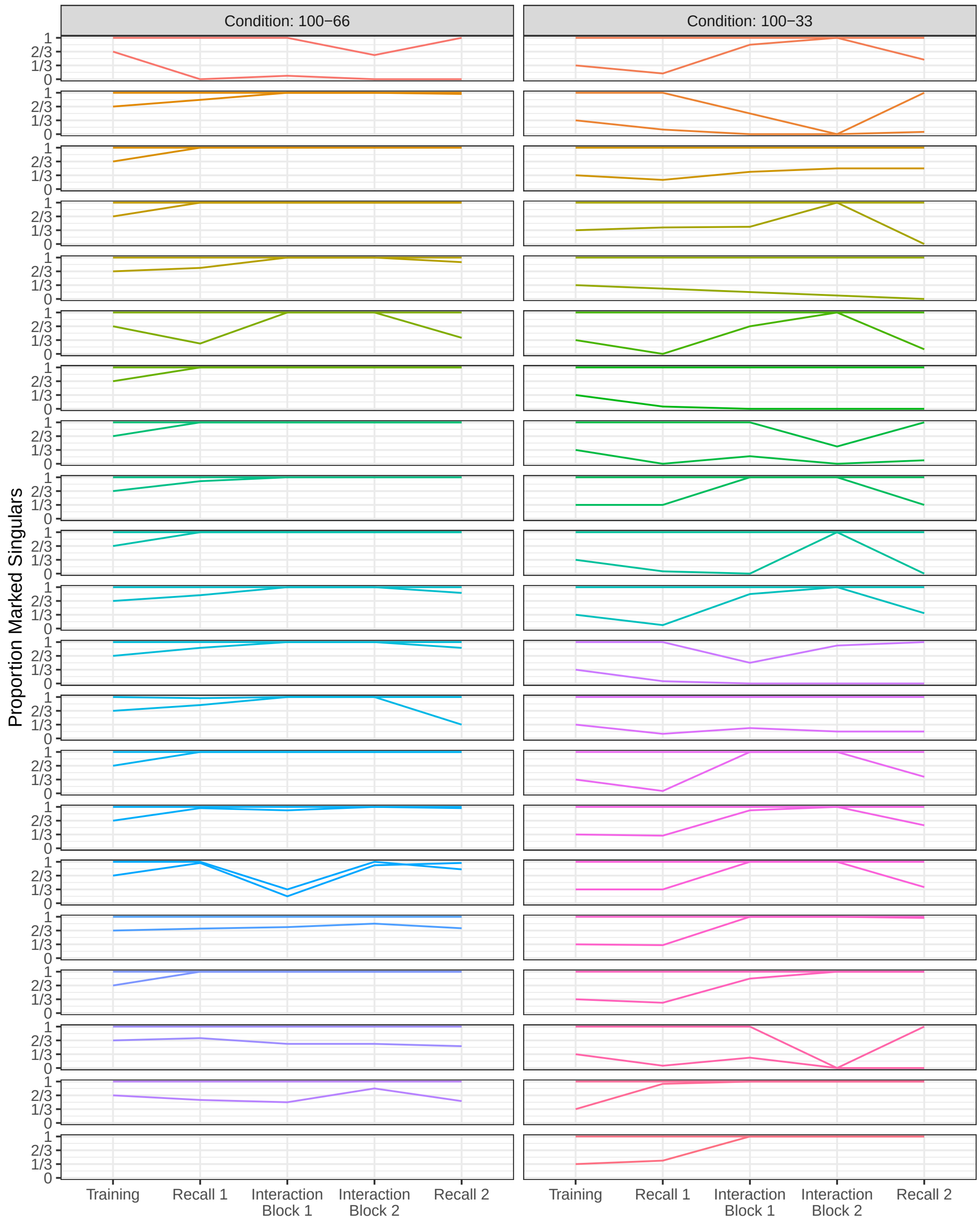

Figure A.2: Experiment 2. Proportion of trials in which the singular was marked, in training (determined by condition), Recall 1, interaction (split by block) and the post-interaction Recall 2, with separate plots showing each pair individually.