COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13435

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13435

Predictability and Variation in Language Are Differentially Affected by Learning and Production

Aislinn Keogh, Simon Kirby, Jennifer Culbertson

Centre for Language Evolution, University of Edinburgh

Received 11 August 2023; received in revised form 1 March 2024; accepted 6 March 2024

Abstract

General principles of human cognition can help to explain why languages are more likely to have certain characteristics than others: structures that are difficult to process or produce will tend to be lost over time. One aspect of cognition that is implicated in language use is working memory—the component of short-term memory used for temporary storage and manipulation of information. In this study, we consider the relationship between working memory and regularization of linguistic variation. Regularization is a well-documented process whereby languages become less variable (on some dimension) over time. This process has been argued to be driven by the behavior of individual language users, but the specific mechanism is not agreed upon. Here, we use an artificial language learning experiment to investigate whether limitations in working memory during either language learning or language production drive regularization behavior. We find that taxing working memory during production results in the loss of all types of variation, but the process by which random variation becomes more predictable is better explained by learning biases. A computational model offers a potential explanation for the production effect using a simple self-priming mechanism.

Keywords: Working memory; Language evolution; Artificial language learning; Regularization; Language production; Urn model

Correspondence should be sent to Aislinn Keogh, Centre for Language Evolution, University of Edinburgh, Edinburgh, EH8 9AD, UK. E-mail: aislinn.keogh@ed.ac.uk

1. Introduction

Language is created in real time: successful processing requires us to rapidly turn complex input into the correct mental representations, while successful production requires us to rapidly turn our mental representations into meaningful output. However, the finite nature of human memory imposes a bottleneck on these processes, shaping the kinds of structures that can persist as languages evolve (Christiansen & Chater, 2016, 2008; Futrell, Mahowald, & Gibson, 2015; Kirby, 1999; MacDonald, 2013). It has long been acknowledged that working memory—the component of short-term memory used for temporary storage and manipulation of information (including linguistic information)—is severely limited in its capacity (Baddeley & Hitch, 1974; Baddeley, 2000; Cowan, 2001; Gobet & Clarkson, 2004; Miller, 1956). Cognitive constraints such as these can help to explain why languages look the way they do: as languages are passed from person to person, properties that make them easier to process or produce are likely to edge out those that place a more significant burden on working memory. Thus, some processes of language change might arise as a result of an interaction between linguistic representations and constraints on memory and other general principles of human cognition (Culbertson & Kirby, 2016).

In this study, we consider the role of working memory limitations in the regularization of linguistic variation. Regularization is a well-documented process of language change whereby a language becomes less variable (on some dimension) over generations. This process has been argued to be driven by individual language learners and users, who produce output that is less variable than their input (Hudson Kam & Newport, 2009). Repeated across many individuals and generations, this behavior is one way in which emerging languages may acquire systematic rules and regularities (Smith & Wonnacott, 2010). For example, nouns in English generally mark plurality with the regular *-(e)s* suffix (e.g., *dog* → *dogs*), but even among irregular nouns there are identifiable, semi-productive patterns (e.g., the vowel change in *mouse* → *mice* and *louse* → *lice*, or null marking in *fish* and *sheep*). Furthermore, while there is considerable variation in the English plural system overall, the choice of form for any given word is generally phonologically or lexically conditioned. By contrast, random variation—where there are no conditioning factors—is rare in natural languages (Givón, 1985), at least in the output of native speakers (Johnson, Shenkman, Newport, & Medin, 1996). Thus, while variation is ubiquitous, it tends to be predictable in some way.

1.1. Regularization of unpredictable variation

There is a wealth of evidence that language users reduce unpredictable variation, both in the lab and in natural language. Children exposed to unpredictable variation in artificial language learning studies tend to regularize at a system-wide level, increasing their use of one variant (usually the form they encountered most frequently in the input) to the exclusion of others (Hudson Kam & Newport, 2005, 2009; Schwab, Lew-Williams, & Goldberg, 2018). This behavior persists even when the most frequent form in the input is not actually very frequent at all (Austin, Schuler, Furlong, & Newport, 2022). Regularization behavior can also be observed in adults, although potentially to a lesser degree or in a narrower range

of circumstances than in children (Culbertson & Newport, 2015; Hudson Kam & Newport, 2009). For example, adults regularize more when the number of alternating variants increases (Ferdinand, Kirby, & Smith, 2019; Hudson Kam & Newport, 2009; Saldana, Smith, Kirby, & Culbertson, 2021), when generalizing to novel contexts (Wonnacott & Newport, 2005), and when attempting to coordinate with other individuals in communicative tasks (Fehér, Ritt, & Smith, 2019; Fehér, Wonnacott, & Smith, 2016; Kamps, Ferdinand, & Kirby, 2014; Perfors, 2016). Furthermore, even when adults maintain variation, they often still regularize at a lower level, making variation more predictable by conditioning it on some aspect of the context like lexical item or grammatical category (Samara, Smith, Brown, & Wonnacott, 2017; Smith & Wonnacott, 2010). And although individual adults may show weaker evidence of regularization than children, this effect may nevertheless be amplified through cultural transmission as small increases in regularity accumulate over generations (Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Smith et al., 2017).

In natural language, regularization of unpredictable variation has been observed in deaf children exposed to inconsistent linguistic input, both in the acquisition of existing signed languages from non-native users (Singleton & Newport, 2004) and in the formation of new signed languages (Senghas, Coppola, Newport, & Supalla, 1997; Senghas & Coppola, 2001). Regularization has also been argued to be at play in the emergence of stable creole languages from highly variable pidgin languages (Aitchison, 1996; Bickerton, 1981; DeGraff, 1999; Siegel, 2007).

1.2. Regularization of predictable variation?

It is less clear whether the cognitive mechanisms driving regularization act as strongly on predictable patterns of variation. In natural language, while there are certainly cases of irregular forms (e.g., *cow* → *kine* in Middle English) shifting to the regular pattern, there is some evidence that *irregularization* is roughly as prevalent a process as regularization, and that the main driver of increased regularity is the introduction of new lexical items (which tend to be regular) rather than the regularization of existing items (Cuskley et al., 2014). Furthermore, regularization is highly frequency-dependent: high-frequency forms tend to exhibit stable irregularity, while lower frequency forms are more likely to regularize (Carroll, Svare, & Salmons, 2013; Cuskley et al., 2017; Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Smith, Ashton, & Sims-Williams, 2023).

Artificial language learning experiments testing the acquisition of conditioned variation also provide somewhat mixed evidence. Although this kind of variation is clearly far more typical of natural language than the unpredictable variation usually targeted by regularization experiments, it is not always learned or reproduced more accurately. When these patterns of variation are only probabilistic, children can struggle, whether conditioning is by linguistic features like syntactic role (Hudson Kam, 2015) or by salient semantic features like natural gender (Schwab et al., 2018). However, children *are* sensitive to certain conditioning cues (especially phonological: Culbertson, Jarvinen, Haggarty, & Smith, 2019; Karmiloff-Smith, 1981; Pérez-Pereira, 1991; Gagliardi & Lidz, 2014) and seem to regularize less (or not at all) when conditioning is deterministic (Austin et al., 2022; Brown, Smith, Samara, &

Wonnacott, 2022; Samara et al., 2017; Wonnacott, 2011). Adults generally have less difficulty acquiring conditioned variation—either probabilistic (Schwab et al., 2018) or deterministic (Austin et al., 2022; Hudson Kam & Newport, 2009)—and often maintain this kind of variation across multiple simulated generations in iterated learning experiments (Smith et al., 2017, Smith et al., 2023; Smith & Wonnacott, 2010). However, as with children, adults' performance varies according to the presence or salience of conditioning cues: neither age group appears to readily acquire arbitrary subclass distinctions (Braine et al., 1990; Culbertson & Wilson, 2013; Frigo & McDonald, 1998; Smith, 1969).

Overall then, there seems to be good reason to suspect that at least certain kinds of conditioned variation will also be regularized—although seemingly to a lesser extent than unpredictable variation.

1.3. What causes regularization?

Whether regularization should target all kinds of variation—or only unpredictable variation—might depend on the underlying cause of the behavior. However, the specific mechanism driving regularization is not agreed upon.

One possibility is that regularization arises from a failure to encode variation during learning (Culbertson, Smolensky, & Wilson, 2013; Hudson Kam & Newport, 2009). In other words, when individuals produce a more regular language than the one they were exposed to, they may be faithfully producing what they remember of their input. On this account, age differences in regularization behavior might be explained by developmental changes in general learning mechanisms; perhaps, by not acquiring the full complexity of their input, children are better able than adults to extract regularities from noise (Hudson Kam & Newport, 2009; Rische & Komarova, 2016). However, tasks that provide a more direct window on individuals' internal representations (e.g., grammaticality judgments or frequency reports) provide evidence that even those who exhibit the most extreme regularization behavior still show awareness of the inconsistencies in their input, including for very complex patterns (Austin et al., 2022; Ferdinand et al., 2019; Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2009; Schwab et al., 2018; Saldana et al., 2021). Furthermore, Perfors (2012) found that requiring participants to attend to a secondary task while they learn an artificial language impaired vocabulary acquisition, but had no effect on the strength of regularization behavior, suggesting that regularization is not an inevitable consequence of imperfect learning.

This suggests that regularization may be primarily a production-side process. However, this still leaves open several possible mechanisms. For example, regularization in production may be driven by specific pragmatic contexts. In line with this, adults seem to regularize more when they understand that the variation in their input is genuinely random (Perfors, 2016), suggesting that when they maintain variation, it is because they think it is meaningful (Clark, 1988). Regularization behavior is also stronger during communicative tasks, either due to accommodation between interlocutors or because individuals strategically remove aspects of the linguistic signal that do not correlate with differences in meaning to maximize communicative success. (Fehér et al., 2016; Fehér et al., 2019). The pragmatic account straightforwardly predicts that unpredictable variation will be regularized, but it is not clear that these

mechanisms would also target predictable variation. Conditioned variation already satisfies language users' expectation that variation in language should be rule-governed (Wonnacott & Newport, 2005), so getting rid of it would not obviously increase communicative success; in fact, failing to observe the rules of the language in this way might even *hinder* communication. Accommodation between interlocutors too would presumably favor the lexically specific rules that both had acquired.

Alternatively, there may be purely cognitive factors that drive regularization in production, such as working memory limitations. One hypothesis which has received some experimental support is that regularization arises from limitations on memory retrieval during language production (Hudson Kam, 2019; Hudson Kam and Chang, 2009). The exact mechanism is unclear, but one possibility is that, when retrieval is difficult, variants that have been produced recently become increasingly accessible for retrieval on subsequent productions through repetition priming (Hudson Kam, 2019; Schwab et al., 2018). These ideas are consistent with models in which language production is not simply a perfect reflection of what has been learned but is also constrained by online demands like ease of retrieval (Goldberg & Ferreira, 2022; MacDonald, 2013). On such an account, we might expect that regularization would target both predictable *and* unpredictable variation since an overall higher frequency form might be more easily retrieved in general, even if specific lexical items had been encountered in different constructions.

Several previous studies suggest that memory retrieval is a factor in driving the regularization of *unpredictable* variation. On the one hand, this hypothesis predicts less regularization when retrieval is less taxing. Indeed, Hudson Kam and Chang (2009) found that adults more closely matched the statistics of their input when the production task was made easier. Similar results have been found with children, who seem to regularize less when the burden of lexical access is eased through the use of English nouns in semi-artificial languages (Samara et al., 2017; Wonnacott, 2011). Another way of getting at the question is to directly interfere with working memory by asking participants to attend to multiple tasks simultaneously. This method aims to disrupt a specific aspect of linguistic working memory—either encoding or retrieval, depending on when it is administered—in order to provide evidence for its involvement. Perfors (2012) performed such a manipulation during learning which, in line with the production-side account, did not result in increased regularization. Hudson Kam (2019) replicated this result with a much more complex language and offered some preliminary evidence that a comparable manipulation during production *may* contribute to increased regularization. Specifically, participants subject to interference during production seemed more likely to regularize on an item-by-item basis (i.e., condition their use of different variants on lexical items).

1.4. The present study

In this paper, we further explore the role of working memory (and memory retrieval) in driving regularization of both predictable and unpredictable variation. In line with Perfors (2012), our goal is to look for evidence of regularization in a simple language which isolates the phenomenon of interest and removes superfluous elements like word order, transitivity,

and negation that are present in the language of Hudson Kam (2019). However, in common with Hudson Kam (2019), we ask whether interfering with working memory during language *production* (rather than learning) leads to regularization. Additionally, we ask whether this production-side mechanism targets predictable variation to the same extent as unpredictable.

To preview, we provide experimental evidence that regularization of both predictable and unpredictable variation does indeed arise under memory load during production. Interestingly, we also find that working memory limitations have some effect on regularization during learning, contrary to previous studies. Finally, we implement a computational model of regularization in production via a simple self-priming mechanism by which a high-frequency variant becomes increasingly accessible for retrieval through repeated production.

2. Experiment

We use a 2×3 between-subjects design to investigate the effect of memory limitations on the regularization of linguistic variation in six experimental conditions. We trained participants on an artificial language exhibiting variation in nominal marking that was either probabilistically lexically conditioned (PREDICTABLE conditions) or random (UNPREDICTABLE conditions). We then tested participants' ability to produce *noun* + *marker* combinations in the language, and their ability to estimate the frequency with which particular *noun* + *marker* combinations had appeared in the input (a measure of learning, following previous work, e.g., Ferdinand et al., 2019). We used an interference task, modeled after the concurrent load tasks used by Perfors (2012) and Hudson Kam (2019), to tax working memory during either learning (LEARNING LOAD conditions) or production (PRODUCTION LOAD conditions); in a third, baseline condition, there was no such task (NO LOAD conditions).

In line with the production-side account of regularization, we predicted that participants would *produce* a more regular language than the one they learned, regardless of the type of variation (predictable or unpredictable). By contrast, we predicted no regularization in participants' frequency estimates. In line with the memory retrieval hypothesis, we predicted that we would see the clearest evidence for reduction of variation when taxing working memory during production. Finally, to test our hypothesis about the relationship between predictable and unpredictable languages, we predicted that the effect of memory limitations during production would be modulated by variation type, with greater regularization of unpredictable languages.

2.1. Methods

The study was approved by the PPLS Ethics Committee at the University of Edinburgh and was pre-registered with the Open Science Foundation (https://osf.io/vqyej).

2.1.1. Participants

We recruited 220 participants via Prolific. Participants were adult, self-reported native English speakers with no known language disorders. They were provided with a downloadable information sheet and gave informed consent to participate. The experiment took around

Table 1
Number of participants per condition submitted to analysis

	Predictable	Unpredictable
No load	29	28
Learning load	30	28
Production load	29	29

Table 2
Distribution of plural markers (P_i) across nouns (N_j) in the two variation conditions

(a) Predictable Input Languages

	N1	N2	N3	N4	N5	N6	Total
P1	7	7	7	7	1	1	30
P2	1	1	1	1	7	7	18
Total	8	8	8	8	8	8	48

(b) Unpredictable Input Languages

	N1	N2	N3	N4	N5	N6	Total
P1	5	5	5	5	5	5	30
P2	3	3	3	3	3	3	18
Total	8	8	8	8	8	8	48

20 minutes to complete ($M = 18.01$, $SD = 8.48$), for which participants were paid £3 (above the UK national minimum wage). Forty-seven participants were excluded for the following pre-registered reasons: self-reporting the use of written notes in an exit questionnaire contrary to instructions (three), data saving errors (one), failing to provide usable data on more than two critical trials (38),[1] and button mashing (five).[2] This left us with data from 173 participants (Table 1).

2.1.2. Materials

The artificial language consisted of orthographically presented labels paired with six images. Each image depicted a pair of animals and was described by a two-word label: one word for the noun and one word indicating plurality (presented in the English frame "Here are two..."). Noun labels were designed to be similar to English onomatopoeia (e.g., "buzzo" for a bee) to ensure that learning of this part of the label would be trivially easy for all participants, regardless of memory load. Nouns were paired with one of two plural markers, both non-English CVC monosyllables ("mej" and "huv"). The mapping of nouns to plural markers varied according to condition (Table 2). In PREDICTABLE conditions, the choice of one plural or the other was probabilistically conditioned on the noun. Four nouns were randomly assigned to one plural marker (the "regulars") and two to the other marker (the "irregulars"). A small amount of noise was then added to this mapping, such that, for n repetitions of a given noun in the training set, that noun appeared with its assigned plural marker $n - 1$ times (87.5%) and once with the other marker (12.5%). This noisy conditioning meant that

participants could regularize without having to produce a description they had never observed. In UNPREDICTABLE conditions, plural markers varied randomly across nouns with no conditioning: all nouns appeared with one marker 62.5% of the time and with the other 37.5% of the time. Both markers appeared with the same overall frequency in the two variation conditions, allowing us to assess the extent to which item-specific patterns affect the tendency to regularize, even when the global language statistics are identical.

2.1.3. Procedure

The experiment was written in JavaScript using the JsPsych library (de Leeuw, 2015) and ran in participants' web browser. Participants were randomly assigned to one of the six conditions at the start of the experiment. The experiment consisted of three phases: training, production, and estimation.

In the training phase, participants were asked to learn the words used to describe the animals. Each of the six images was shown eight times for a total of 48 trials. The order of presentation was randomized. On each training trial, an image was presented for 1000 ms and then a description of the form "Here are two *noun* + *plural*" appeared below the image. The image and description disappeared after 3000 ms and participants clicked a "continue" button to advance to the next trial.

In the production phase, participants were asked to produce descriptions for the same set of stimuli. Again, each of the six images was shown eight times for a total of 48 trials.[3] On each production trial, participants saw an image and a partial description, consisting of an English frame and two gaps for the artificial words: "Here are two _____ _____". They were asked to fill in the gaps by clicking two buttons from an array consisting of all nouns and plural markers in the language. This multiple-choice production task is intended to simulate the process of a fluent speaker selecting words from a stably represented mental lexicon. It allows us to observe the effects of online demands in production while minimizing the possibility that participants' choice of words is driven by incomplete learning.[4] Buttons were blocked into nouns (on the left) and plural markers (on the right), with the order of buttons randomized within each block and a clear gap between blocks. However, participants were not forced to click one button from the first block and one from the second. No feedback was provided; participants simply saw the gaps filled with whichever words they had selected. The full label they had assembled was displayed for 1000 ms before they advanced to the next trial.

Finally, in the estimation phase, participants were asked to estimate how often they had seen each noun with each plural marker in training. All six images appeared in a random order on one page, each accompanied by a continuous slider over percentages. All sliders started in the middle, and participants were required to move every slider before they could advance. Each slider had three labels: "always *P1*" at 0%, "equal *P1/P2*" at 50%, and "always *P2*" at 100%. The assignment of plural markers to the two ends of the slider was randomized for each participant, but identical for all sliders.

In LEARNING LOAD and PRODUCTION LOAD conditions, participants were told that we were interested in how well people can learn or produce (respectively) a new language when the task is difficult, so they would also be asked to memorize and recall short sequences of numbers alongside the main task. They were told that they would be given feedback throughout

on their performance on this task. The aim was to occupy participants' conscious attention with the secondary task to disrupt the part of working memory they would otherwise have devoted to the linguistic task. The task was sandwiched around (i.e., concurrent with) each trial in either the training phase (LEARNING LOAD) or production phase (PRODUCTION LOAD). First, a pseudorandom sequence of three digits was displayed for 2500 ms and participants were asked to memorize the numbers in order. A new sequence was generated on each trial by sampling the set of digits 0–9 without replacement, with the constraint that each digit n was never neighbored on either side by $n + 1$ or $n - 1$, preventing any obvious patterns appearing in the sequences that might have made them easier to remember. Participants then completed the main training or production trial. Immediately following this, participants were asked to retype the numbers they had just memorized, in order. They were given feedback on the number of digits they had recalled in the correct position and how long they had taken to respond, to encourage both speed and accuracy.

A schematic of the experimental procedure for the PRODUCTION LOAD conditions is given in Fig. 1.

2.1.4. Analysis

We take an information theoretic approach (Shannon, 1948) to quantifying variation and regularization (following, e.g., Ferdinand et al., 2019; Perfors, 2016; Samara et al., 2017; Smith & Wonnacott, 2010). This analytic approach is sensitive even to small changes in frequency distributions, regardless of whether those changes are in the direction predicted by the input (i.e., even if participants regularize with the minority variant[5]). We report three specific measures below: entropy, conditional entropy, and mutual information (MI). The first two measures were pre-registered, the third is an addition which we explain below.

Entropy: The total amount of variability in a plural marking system is captured by the entropy of the frequency distribution of plural markers across the language. Taking plural marking as a discrete random variable V with possible variants $v_1 \ldots v_n$ which occur with probability $p(v_1) \ldots p(v_n)$, the entropy of a language is given as

$$H(V) = -\sum_{v_i \in V} p(v_i) \log_2 p(v_i).$$

More skewed distributions (i.e., languages in which one plural marker is used more frequently) exhibit lower entropy. A maximally regular language (with only one plural marker) would score 0, while a maximally irregular language (where both markers appear 50% of the time) would score 1. Since the frequency distribution of plural markers across the input languages in both PREDICTABLE and UNPREDICTABLE conditions is identical, the languages are matched for entropy (0.95 bits).

Conditional entropy: The predictability of a plural marking system can be measured by considering how variable individual nouns are: a language where each noun only uses one plural marker is more predictable than one where nouns can take any marker. The average variability of individual nouns in a language is captured by the conditional entropy of the

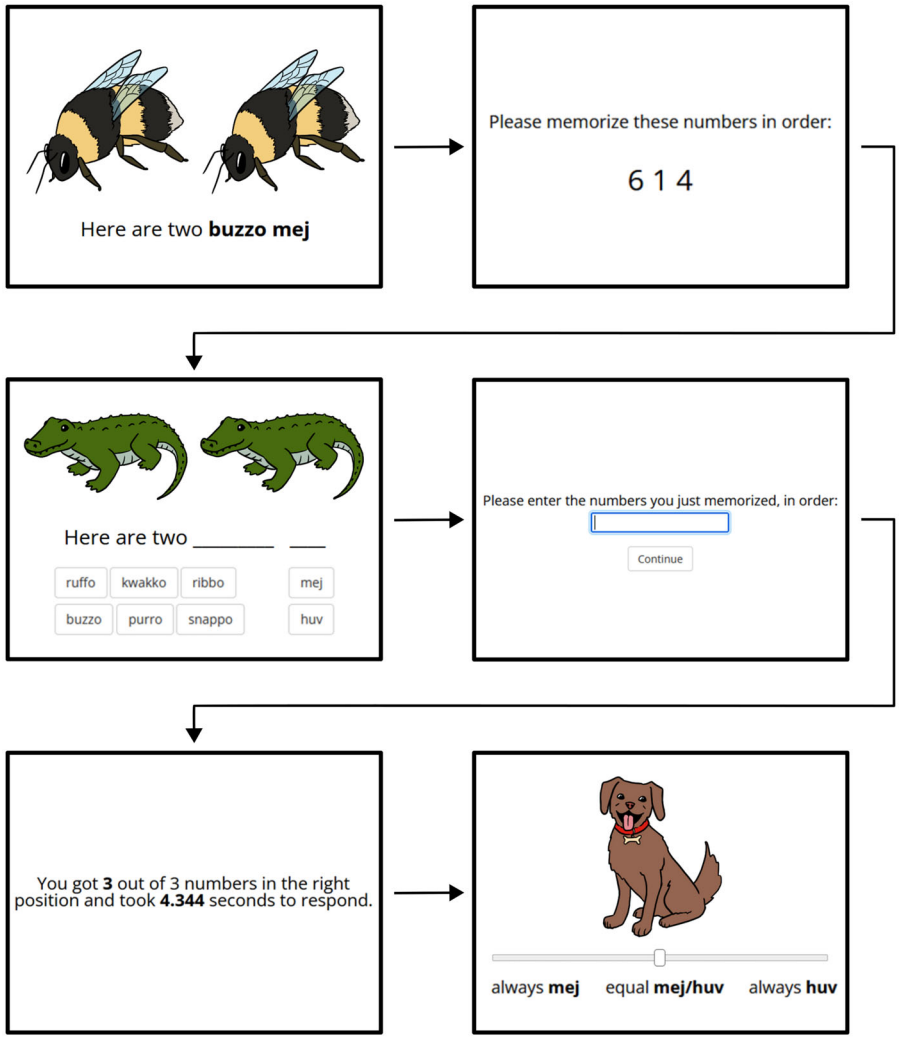

Fig. 1. Schematic of the experiment: PRODUCTION LOAD condition. Top to bottom, following arrows: training trial, digit sequence presentation, production trial, digit sequence recall, feedback, and estimation trial. Participants in LEARNING LOAD conditions would instead have seen the digit sequence presentation and recall trials sandwiched around each training trial. Participants in NO LOAD conditions would not have seen these digit sequence trials.

frequency distribution of plural markers, given the noun being marked. Given a set of variants V (plural markers) and a set of contexts in which these variants appear C (nouns), the conditional entropy of a language is given as

$$H(V|C) = -\sum_{c_j \in C} p(c_j) \sum_{v_i \in V} p(v_i|c_j) \log_2 p(v_i|cj).$$

The variability of individual nouns in UNPREDICTABLE input languages mirrors that of the language as a whole, so entropy and conditional entropy are matched for these languages (0.95 bits on both measures). On the other hand, PREDICTABLE input languages have lower conditional entropy since individual nouns in these languages are less variable than the language as a whole (0.54 bits).

Mutual information: When either entropy or conditional entropy decreases, we can infer that the language has become more regular in some sense. However, here we would like to distinguish between regularization at the lexical level (i.e., a given plural marker used more with a particular noun) and regularization across the language as a whole (i.e., a given plural marker used more often overall). Conditional entropy does not allow us to do this since it is affected by overall entropy: when a language becomes less variable overall, the choice of plural marker necessarily becomes more predictable. We, therefore, added a third measure to our set of pre-registered variables: mutual information (MI). MI is the difference between the two entropy measures[6] and allows us to isolate the amount of predictability that is specifically explained by lexical conditioning. MI of 0 indicates a complete absence of lexical conditioning; this is the case both when there is no variability (since there is nothing to condition here), and when the variability of individual nouns mirrors that of the language overall (as in the UNPREDICTABLE input languages). MI of 1 would indicate that the language as a whole is maximally variable (i.e., the two plural markers are equally frequent overall), but each noun is perfectly non-variable. PREDICTABLE input languages here score 0.41, reflecting the presence of imperfect conditioning in a skewed overall frequency distribution.[7]

2.2. Experiment results

We analyzed the data in R (R Core Team, 2022). Each of the measures described in Section 2.1.4 was calculated for the languages participants were trained on, the languages they produced, and the languages described by their estimates. We investigate regularization as a function of learning by comparing participants' estimates to their training data. We investigate regularization as a function of production by comparing participants' productions to their training data and to their estimates. The dependent variable in all analyses is, therefore, the *change* in the given measure. We define regularization as a reliable *decrease* in entropy or a reliable *increase* in MI. Plots in this section show population-level data; individual-level data are available in Appendix A.

2.2.1. Pre-requisites

In order to test the hypotheses of interest, it is crucial that we first rule out the possibility that any differences between conditions are driven by differences in vocabulary learning or in performance on the interference task. The following mixed effects models were generated using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) and include fixed effects of variation type and memory load, and their interaction, as well as by-participant and by-item random intercepts.

Performance on the interference task was close to ceiling across conditions (overall, $M = 2.84$, $SD = 0.57$). Since the distribution of scores is very left-skewed, we take as our dependent variable the *error rate* (calculated as 3—the number of correct digits), which approximates a Poisson distribution. We performed a mixed effects Poisson regression predicting the error rate by condition. Model comparison revealed that neither variation type ($\chi^2(2, 6) = 0.258$, $p = .879$) nor memory load ($\chi^2(2, 6) = 2.462$, $p = .292$) were significant predictors of performance. These results indicate that participants in all conditions were attending equally well to this task. Furthermore, the level of performance indicates that participants took the task seriously; we can, therefore, be confident that participants did not focus on the main task to the exclusion of the interference task, which would obscure any possible effects in the load conditions.

Noun learning was also close to ceiling across conditions (overall, $M = 0.97$, $SD = 0.17$). We performed a mixed effects logistic regression predicting the log-likelihood of a correct response by condition. Model comparison revealed that neither variation type ($\chi^2(3, 8) = 1.278$, $p = .734$) nor memory load ($\chi^2(4, 8) = 2.719$, $p = .606$) were significant predictors of noun learning. These results indicate that participants in all conditions learned the lexicon equally well.

In summary, any differences in regularization we see across conditions are not due to accidental differences in performance on the memory load task or noun learning.

2.2.2. Main analysis

Inspection of the models specified in our pre-registration revealed that residuals were significantly non-normally distributed (confirmed by Shapiro–Wilk tests) and had non-constant variance over groups (confirmed by Breusch–Pagan tests for heteroscedasticity). Since our data did not meet the assumptions for a linear modeling analysis, the analyses we present here instead evaluate our pre-registered predictions using a simulation-based approach.[8]

Our null hypothesis is that participants' responses reflect a probability-matching strategy (e.g., Estes, 1976; Gardner, 1957; Hudson Kam & Newport, 2005). To determine how much we can expect entropy and MI to change under this strategy, we simulate participants who produce the majority marker for any given noun on any given trial with a probability equal to its frequency in the input. We generate 10,000 runs of 30 such participants and calculate the mean of each run. This gives us a distribution of expected means under the null hypothesis against which we can z-score our real by-condition means. A z-score of < -1.96 indicates a reliable decrease in entropy, while a z-score of > 1.96 indicates a reliable increase in MI.[9]

To identify main effects of our predictors, we take a permutation-based approach. The null hypothesis is that different conditions do not give rise to substantially different behavior. We can generate data that meets this assumption by randomly shuffling the labels for one predictor in our real data. For example, to test for a main effect of variation type, we shuffle the column containing the PREDICTABLE/UNPREDICTABLE labels, thus breaking the association between each data point and its condition label. We carry out this shuffling 10,000 times, calculating the difference between condition means (in the example case, between the mean of all PREDICTABLE and all UNPREDICTABLE conditions) for each run, to give us a distribution of expected differences between conditions under the null hypothesis, against which we can

z-score our real difference.[10] A z-score of > 1.96 indicates that the observed difference between conditions is reliably greater than would be expected by chance.

This permutation analysis also allows us to identify interactions between predictors. The null hypothesis here is that the difference between the levels of one predictor is the same across the levels of the other predictor, that is, the effect of memory load does not depend on variation type or vice versa. Again, we can generate data that meets this assumption by randomly shuffling the labels for one predictor in our real data. For example, to test whether the effect of the PRODUCTION LOAD manipulation differs between variation types, we first shuffle the column containing the PREDICTABLE/UNPREDICTABLE labels then calculate the difference between the PRODUCTION LOAD condition and other memory load conditions (collapsed) separately for the PREDICTABLE and UNPREDICTABLE conditions, and finally calculate the difference between these differences. We carry out this shuffling 10,000 times to generate a distribution of expected differences in differences under the null hypothesis, against which we can z-score our real difference in differences. A z-score of > 1.96 indicates that the observed difference in differences is reliably greater than would be expected by chance.

We can also calculate p-values for all reported statistics by counting the number of values in the relevant null distribution that are as or more extreme than our observed value and dividing this by the number of runs (10,000). Due to the finite nature of the sample, this sometimes gives a value of exactly 0 or 1; in this case, we report $p < .001$ or $p > .999$.

Regularization during learning: We predicted that participants across conditions would show no evidence of having learned a more regular language than the one they were trained on. The estimation task results allow us to assess this prediction. The comparison of interest is thus between the languages participants were trained on and the ones described by their estimates.

Fig. 2a shows the change in entropy. In line with our prediction, we found no reliable decrease in entropy: no condition mean falls below the lower tail of the corresponding null distribution. However, as Fig. 2a shows, there was an increase in MI between the languages participants in UNPREDICTABLE conditions were trained on and the ones described by their estimates: the mean of each of these conditions is well above the null distribution. Permutation analysis confirms a main effect of variation type ($Z = 5.498$, $p < .001$).

To summarize, these results show, in line with our prediction, that the learning process does not drive regularization at a system-wide level: participants are able to encode the overall frequency of different variants in their input. However, we do see evidence of a learning bias for regularization at the lexical level, with learners in the UNPREDICTABLE conditions inferring a pattern of conditioning when no such pattern exists in their input.

Regularization during production: Before analyzing participants' production data, we got rid of trials where the label produced was of an invalid form (i.e., anything other than *noun* + *plural*) or where the noun was incorrect.[11]

Recall that we predicted that taxing working memory during production would lead to greater regularization behavior. We also predicted that we would see greater regularization of unpredictable languages and that this factor would modulate the size of the effect of memory

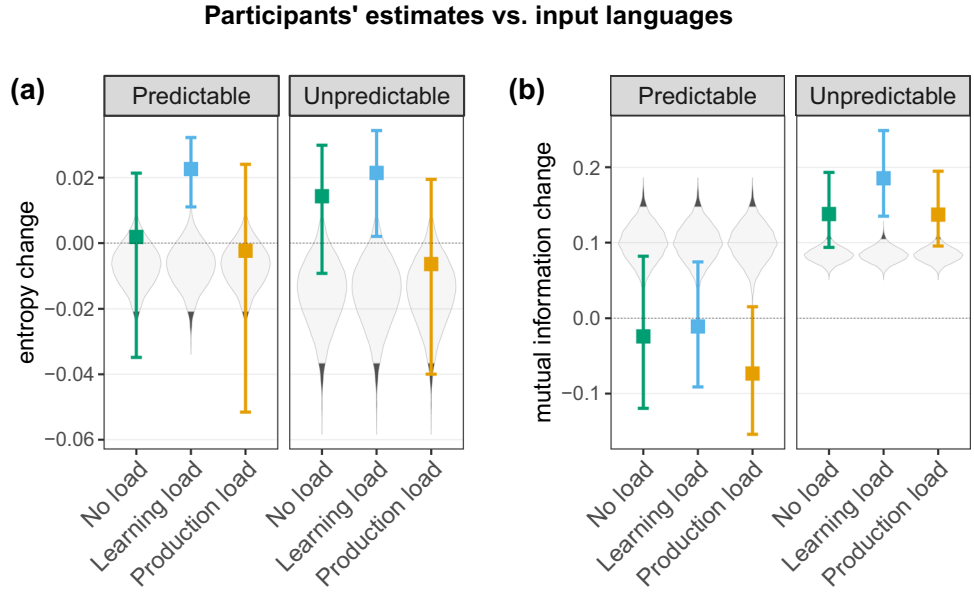

Fig. 2. Change in entropy (left) and mutual information (right) between the languages participants were trained on and the ones described by their estimates, by condition. Points represent condition means; error bars represent bootstrapped 95% confidence intervals over the mean. Violins show the distribution of expected means under the null hypothesis of probability-matching; regularization is indicated by means below the lower tail (entropy) or above the upper tail (MI) of these distributions. There is no reliable decrease in entropy in any condition, indicating that participants did not underestimate the total amount of variation in their input. However, there is a reliable increase in MI between the languages participants in UNPREDICTABLE conditions were trained on and the ones described by their estimates, indicating that participants in these conditions overestimated the degree of lexical conditioning present in their input.

limitations. To assess these predictions, the comparison of interest is between the languages participants were trained on and the ones they produced.

Fig. 3a shows the change in entropy. In line with the first part of our prediction, the only place we see a reliable drop-in entropy is the PRODUCTION LOAD conditions: the means of these conditions (and no others) are both below the lower tail of the null distributions. Permutation analysis confirms a main effect of memory load, with greater entropy drop in PRODUCTION LOAD conditions than other memory load conditions ($Z = -3.034$, $p = .001$). Contrary to our prediction, permutation analysis reveals no main effect of variation type ($Z = 0.620$, $p = .733$). Although, descriptively, entropy does drop more in the UNPREDICTABLE/PRODUCTION LOAD condition ($M = -0.118$) than in the PREDICTABLE/PRODUCTION LOAD condition ($M = -0.060$), we find no statistical evidence that the effect of the production load manipulation is stronger in the UNPREDICTABLE condition ($Z = 1.092$, $p = .856$). In other words, there is no reliable interaction between variation type and memory load.

As shown in Fig. 3b, we observed an increase in MI across all conditions apart from PREDICTABLE/NO LOAD ($Z = 1.506$, $p = .065$) and PREDICTABLE/PRODUCTION LOAD ($Z = -2.650$, $p = .996$). On this measure, our data, therefore, suggest that there is a general

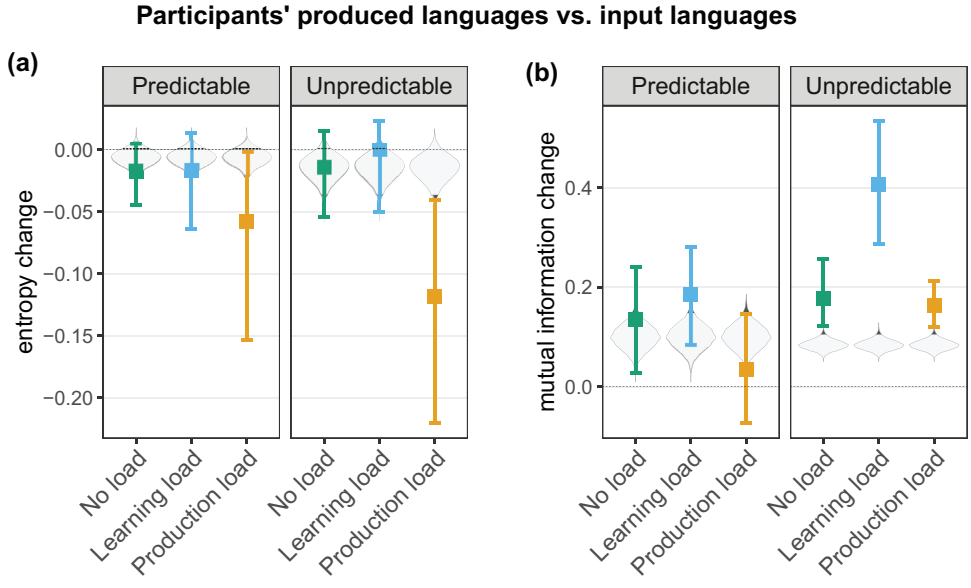

Fig. 3. Change in entropy (left) and mutual information (right) between the languages participants were trained on and the ones they produced, by condition. Points represent condition means; error bars represent bootstrapped 95% confidence intervals over the mean. Violins show the distribution of expected means under the null hypothesis of probability matching; regularization is indicated by means below the lower tail (entropy) or above the upper tail (MI) of these distributions. Entropy decreases only in PRODUCTION LOAD conditions, indicating that taxing working memory during production increases participants' tendency to over-produce one variant relative to its frequency in the input. MI, on the other hand, increases in all but the PREDICTABLE/NO LOAD and PREDICTABLE/PRODUCTION conditions, and especially so in the UNPREDICTABLE/LEARNING LOAD condition. This seems to reflect a general preference to produce lexically conditioned variation, amplified by memory limitations during learning.

tendency to introduce or boost lexical conditioning, not arising from the same memory mechanism that leads to entropy drop. In line with our prediction, permutation analysis confirms a main effect of variation type, with a greater increase in MI in UNPREDICTABLE conditions ($Z = 2.999$, $p = .002$).[12] Permutation analysis also reveals a main effect of memory load. However, as suggested by Fig. 3b, this is in the opposite direction than predicted: MI increases *less* in PRODUCTION LOAD conditions than other memory load conditions ($Z = -2.745$, $p = .002$). Since inspection of the means suggests that MI actually increased more in LEARNING LOAD conditions, we carried out an exploratory analysis by collapsing NO LOAD and PRODUCTION LOAD conditions together. Permutation analysis on this coding scheme supports the notion that MI increases significantly more in LEARNING LOAD conditions than others ($Z = 3.596$, $p < .001$), suggesting that the preference for lexical conditioning is amplified by memory limitations during *learning*. The interaction analysis we ran for the entropy data is clearly not warranted by the MI data since the main effect does not go in the predicted direction. We carried out a further exploratory analysis comparing LEARNING LOAD conditions to other memory load conditions, but this analysis revealed no reliable interaction between variation type and memory load ($Z = -1.407$, $p = .081$).

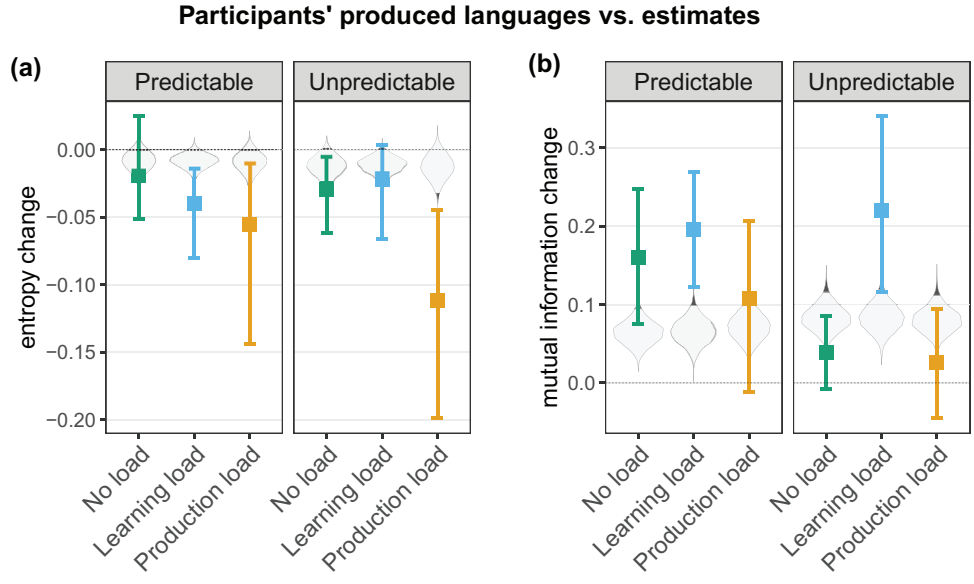

Fig. 4. Change in entropy (left) and mutual information (right) between the languages described by participants' estimates and the ones they produced, by condition. Points represent condition means; error bars represent bootstrapped 95% confidence intervals over the mean. Violins show the distribution of expected means under the null hypothesis of probability matching; regularization is indicated by means below the lower tail (entropy) or above the upper tail (MI) of these distributions. The same memory manipulation that drives regularization behavior during production also predicts how much more regular participants are in production than in their estimates in terms of entropy change. Participants produce a more deterministic pattern of conditioning than the one described by their estimates in the majority of conditions; however, in the UNPREDICTABLE/NO LOAD and UNPREDICTABLE/PRODUCTION LOAD conditions, learning effects account for all the increase in MI seen in production.

We also predicted that participants in all conditions would produce a more regular language than the one described by their estimates. This pattern is what was found by Ferdinand et al. (2019), who use it to argue that regularization is driven by production-side biases. In addition, we predicted that the same factors that drive regularization behavior during production should explain differences in regularity between participants' productions and their estimates. Taken together, we thus predicted that differences across conditions in the regularity of productions compared to input would be replicated when comparing productions to estimates.[13] In other words, when plotting the change in entropy and MI by condition, we would expect to see similar patterns for the production-input comparison and the production-estimate comparison.

Fig. 4a shows the difference in entropy. On this measure, participants were more regular in production than in their estimates in all conditions except PREDICTABLE/NO LOAD ($Z = -1.385$, $p = 0.90$) and UNPREDICTABLE/LEARNING LOAD ($Z = -1.579$, $p = .067$). In line with our prediction, the same memory manipulation that drives regularization behavior during production also predicts how much more regular participants are in production than in their estimates: permutation analysis confirms a main effect of memory load, with greater entropy

drop in PRODUCTION LOAD conditions than other memory load conditions ($Z = -2.527$, $p = .007$). As in the production-input comparison, permutation analysis shows no main effect of variation type ($Z = 0.796$, $p = .218$), and no interaction between variation type and memory load ($Z = 1.075$, $p = .150$).

Fig. 4b shows the difference in MI between participants' productions and their estimates. On this measure, participants were more regular in production than in their estimates in all conditions except UNPREDICTABLE/NO LOAD ($Z = -2.700$, $p = .997$) and UNPREDICTABLE/PRODUCTION LOAD ($Z = -3.394$, $p = .999$), suggesting that the increase in MI seen in participants' productions is accounted for by learning effects in these conditions. Unlike in the production-input comparison, permutation analysis shows no main effect of variation type ($Z = 1.635$, $p = .051$). However, as in the production-input comparison, permutation analysis reveals a main effect of memory load in the opposite direction than predicted: MI increases less in PRODUCTION LOAD conditions than others ($Z = -2.251$, $p = .011$). Exploratory analysis comparing LEARNING LOAD conditions to other memory load conditions (collapsed) supports the notion that MI increases significantly more in LEARNING LOAD conditions than others ($Z = 3.102$, $p < .001$). Again, the interaction analysis we ran for the entropy data is clearly not warranted by the MI data since the main effect does not go in the predicted direction. We carried out a further exploratory analysis comparing LEARNING LOAD conditions to other memory load conditions (collapsed), but this analysis revealed no reliable interaction between variation type and memory load ($Z = 1.551$, $p = .060$).

To summarize, these results show, in line with our prediction, that reduction of overall variability is driven by memory limitations during language *production*. By contrast, lexical conditioning is boosted relative to the input almost across the board, and this tendency is even more pronounced when memory is taxed during *learning*.

Fig. 5 shows an example of one participant's behavior across the experiment. This participant was in the UNPREDICTABLE/LEARNING LOAD condition, so they were trained on a language with a 62.5/37.5 split between the two plural markers for every noun. Their estimates describe a very different language: one where four nouns *only* appear with the majority marker,[14] one noun *only* appears with the minority marker, and the remaining noun has a roughly 50/50 split between the two plurals. This language has entropy of 0.82 (compared to the input entropy of 0.95) and MI of 0.64 (compared to the input MI of 0). The language they produced was even more regular than their estimates in terms of lexical conditioning, with MI of 0.85, but almost identical to the input in terms of the overall frequency distribution of plural markers, with entropy of 0.94.

2.3. Discussion

In this experiment, we investigated whether working memory limitations during production drive regularization of both predictable (conditioned) and unpredictable (random) variation. In line with this hypothesis, we found evidence for a reduction in both types of variation when memory was taxed during production. As in previous research (e.g., Ferdinand et al., 2019; Hudson Kam & Newport, 2009; Saldana et al., 2021; Schwab et al., 2018), this effect was not driven by learners failing to accurately encode the overall frequency of different variants

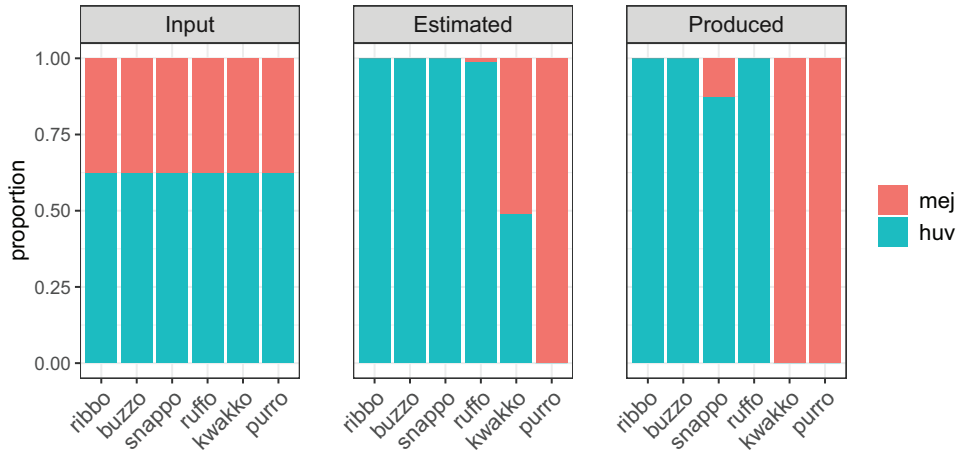

Fig. 5. Example language estimated (middle) and produced (right) by one participant in the UNPREDICTABLE/LEARNING LOAD condition, relative to the input (left). The participants' estimates indicate that they learned a pattern of lexical conditioning that was not present in the input; they then made this pattern even more deterministic in production.

in their input. Importantly though, our results do not support an exclusively production-side account of regularization. In particular, we found evidence for an increase in lexical conditioning during both learning and production. In other words, a bias to reduce variability by increasing conditioning affects both language users' inferences during learning and their (implicit) decisions during production.

Although we saw a reduction in overall variation when taxing memory during production (a drop in entropy), we also observed a different kind of regularization in this experiment: an increase in lexical conditioning. However, this effect was *not* driven by memory load during production and was, if anything, amplified by taxing memory during *learning*. This suggests that working memory limitations during language production can account for regularization at the system-wide level but not at the lexical level. In other words, language users might overproduce particular variants (relative to their frequency in the input) as a result of limitations on memory retrieval, but this is not the mechanism by which variation becomes lexically conditioned. This begs the question: What assumptions do we need to make about memory retrieval processes in order to explain this discrepancy? In other words, *how* do limitations on working memory during language production give rise to some properties of regularity but not others? We turn to this question in the next section.

3. A model of production-side regularization

Historically, computational work has sought to explain regularization as a function of *learning* biases (e.g., Culbertson et al., 2013; Perfors, 2012; Ramscar & Gitcho, 2007; Ramscar & Yarlett, 2007; Reali & Griffiths, 2009; Rische & Komarova, 2016). However, in our

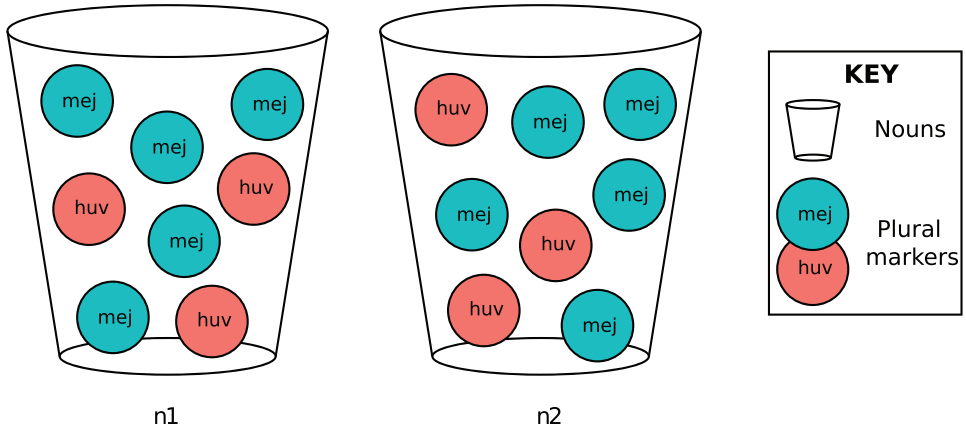

Fig. 6. An urn model conceptualization of nominal plural marking. Plural markers are represented as balls in urns (nouns). When agents encounter a noun n_i, they produce a plural marker by choosing a ball at random from the associated urn U_{n_i}. In this case, the agent would produce "mej" with a probability of 0.625 for either noun.

experiment, we found that working memory limitations operating during language *production* were a reliable predictor of regularization behavior. Furthermore, learning data (from the estimation task) did not reveal any prior bias for the kind of regularity we saw emerging in PRODUCTION LOAD conditions, that is, an overall loss of one variant in favor of another.

Previous work has suggested that the mechanism by which memory constraints result in regularization is overretrieval of a more accessible form (Goldberg & Ferreira, 2022; Hudson Kam and Chang, 2009; Marcus et al., 1992). More specifically, recent research (Hudson Kam, 2019; Schwab et al., 2018) has speculated that a kind of repetition priming might drive increased accessibility of forms that have been produced more recently. Here, we implement this mechanism in a simple "urn" model (Hintzman, 1986; Nosofsky, 1986; Spike, Stadler, Kirby, & Smith, 2017; Walsh, Möbius, Wade, & Schütze, 2010). We show that such a model can capture the entropy decrease in our experimental PRODUCTION LOAD conditions by means of a production process that causes one variant to be retrieved more than would be predicted by its frequency in the input.

3.1. Details of the model

Urn models represent the object of interest (here, plural markers) as balls in an urn or set of urns (here, nouns), where different variants correspond to different colored balls (Fig. 6). In the basic urn model, an agent draws a ball randomly from an urn and observes its color, places it back in the urn, and then repeats the selection process. Here, we model the memory load effect as a simple self-priming mechanism using a Pólya urn model (see Mahmoud, 2008, for an overview). In a Pólya urn model, k additional balls of the same color are added to the urn after each draw. In this way, the probability of producing a particular variant depends not only on that variant's frequency in the input but also on the frequency with which it has

already been produced; observed values become more likely to be observed again. In other words, variants that are produced more become even more accessible for retrieval in future trials than would be predicted by the input statistics alone. Note that this process does not inevitably favor the variant that had a higher frequency in the input: as long as an urn contains both variants, it is always possible that the lower frequency one will be chosen on the first trial and then boosted by the priming mechanism.

The population is a set of agents A who each learn a language L. Here, the language is a set of nouns $\{n_1, \ldots, n_6\} \in N$, each with an associated urn U_{n_i} containing plural markers from the set $\{p_1, p_2\} \in P$. Since participants in the real experiment did not always learn the input languages perfectly, we used the languages described by participants' estimates as the input to our agents. In this way, we can model the effect of production mechanisms *after* taking learning effects into account.

Each agent a completes 48 production trials—eight for each noun (as in the real experiment). On each trial, the agent encounters a random noun n_i and produces a plural marker for that noun by sampling the corresponding urn U_{n_i}. L is then updated according to the parameters described in the next section.

3.1.1. Parameters

In order to find a model that would provide the best fit to the experiment data, we consider all combinations of the following parameter settings for both PREDICTABLE and UNPREDICTABLE input languages. These parameters are intended to spell out the details of how self-priming through repeated production can give rise to regularization, and where this behavior comes from—both at an individual and population level.

Priming scope: One possibility is that priming is context-sensitive: the variant that was most recently produced for a given noun is more likely to be produced the next time *that noun* is encountered. Alternatively, priming could be context-agnostic: the variant that was produced on trial t_i is more likely to be produced on trial t_{i+1}, regardless of which nouns are encountered on those two trials. The *priming scope* parameter, therefore, has three possible values: within nouns, between nouns, or both.

Priming strength: Although we did observe regularization at a population level in our experimental PRODUCTION LOAD conditions, there was substantial variation in the extent to which individual participants showed this effect. We, therefore, wanted to allow agents in the model to differ systematically from each other in the same way. To do so, we randomly select a value of k for each agent: the number of additional balls they add to the relevant urns after each draw. Thus, the strength of the priming mechanism is a property of individuals, not a property of populations. We allow k to range between 0 and 8: at most, agents can add the same number of balls as were in the urn to start with, but it is possible for them to add none (and they can never take any away). Two parameters control the way k is selected.

First, we model the distribution of k in the population according to one of three distributions from the beta-binomial family: uniform ($\alpha = \beta = 1$), normal-like ($\alpha = \beta > 1$), or u-shaped ($\alpha < 1, \beta < 1$). These distributions capture different types of populations. In the uniform

distribution, all values of k are equally likely; in such a population, there is no concept of a "typical" agent. In the normal-like distribution, values around the mean are the most likely and extreme values (in either direction) are less likely. In the u-shaped distribution, extreme values are *more* likely. Specifically, we parameterize this distribution such that the most likely value is 0, the maximum value (given the range) is about half as likely, and values in the middle are considerably less likely. Concretely, approximately 90% of agents will use a value of k at one of the two extremes of the range.

Second, we consider all mean values of k in the set $\{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ for each distribution.[15] We use this value to set the upper bound on the range of allowable values.[16]

We sample k according to the following procedure:

$$k \leftarrow \begin{cases} \text{random.betabinom } (n = 2m,\ \alpha = 1,\ \beta = 1), & \text{if } d == \text{“uniform”} \\ \text{random.betabinom } (n = 2m,\ \alpha = 100,\ \beta = 100), & \text{if } d == \text{“normal-like”} \\ \text{random.betabinom } (n = 2m,\ \alpha = 0.05,\ \beta = 0.1), & \text{if } d == \text{“u-shaped”} \end{cases}$$

where m is the *mean priming strength* and d is the *population distribution*.

Forgetting: In the basic Pólya urn model, the number of balls increases at every time step when $k > 0$. We do not consider this situation here, since it would have the somewhat implausible effect that agents who are most affected by the self-priming mechanism (i.e., those with the most severely limited working memory) would also end up storing the largest number of data points in memory. An alternative model is one where the amount of data remains constant through the deletion of k balls from each urn for k that are added. We consider two deletion methods: either k balls are randomly removed from the urn, or deletion always targets the k oldest balls. The *forgetting* parameter, therefore, has two possible values: random or oldest. Importantly, forgetting never preferentially targets the low-frequency variant (a condition that was proposed to be essential in modeling of regularization during learning by Perfors, 2012).

3.1.2. Analysis

Each model is a unique combination of parameter settings. We ran 100 simulated experiments with each model, each consisting of the same number of agents in the PREDICTABLE and UNPREDICTABLE conditions as in the corresponding PRODUCTION LOAD conditions in the real experiment. For each experiment, we calculated the mean change in entropy and MI (relative to the input) by condition and obtained a 95% confidence interval around these means through bootstrapping. We then averaged over the 100 experiments. To determine which model provides the best fit to the experiment data, we compared these simulated means and confidence intervals to the corresponding means and confidence intervals of the PRODUCTION LOAD conditions in the real experiment. Each model received a divergence score, which captures the average absolute difference between the real and simulated means and confidence intervals across conditions; lower scores indicate that the data generated by that model are more similar to the real data.

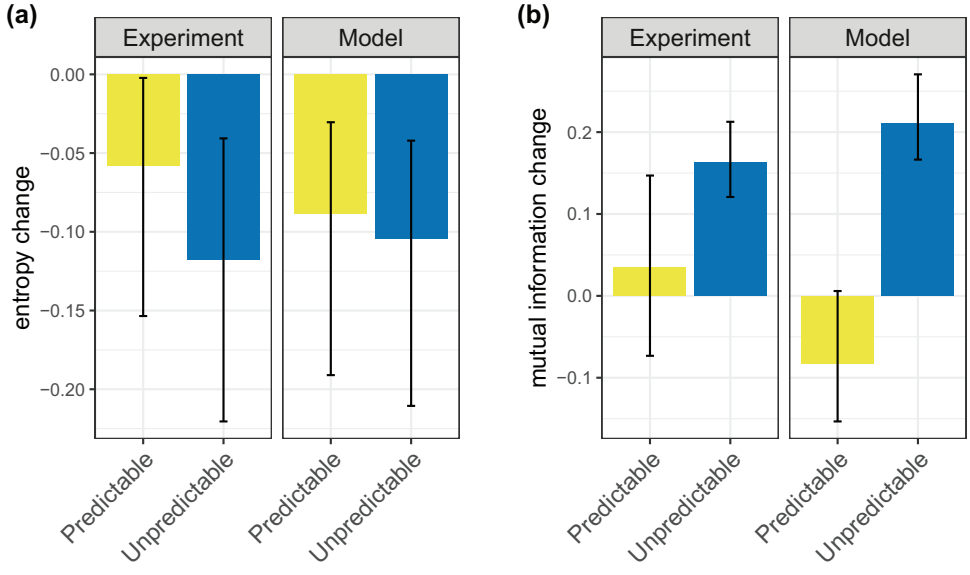

Fig. 7. Change in entropy (left) and mutual information (right) between input language and production output for participants in the experimental PRODUCTION LOAD conditions and agents in the best-fit model. Parameter settings were as follows: priming both within and between nouns, k (the priming strength parameter) drawn from a u-shaped distribution with mean 2.0, and random forgetting.

3.2. Model results

Fig. 7 shows the data generated by the model that provided the best fit to the experiment data overall. This model had priming both within and between nouns. The priming strength parameter k was drawn from a u-shaped distribution with median 2.0, that is, k could take any value in the set $\{0, 1, 2, 3, 4\}$, but extreme values were more likely. Balls were randomly selected for deletion after new ones were added. Further details of the performance of different parameter settings are available in Appendix B.

The inter-agent variation that arises by sampling k from some distribution on an agent-by-agent basis is a demonstrably key component of these models. Fig. 8 shows the entropy results for two models where all agents use the same value of k: either 1 (the lowest possible non-zero value) or 4 (the highest possible value in the distribution used by the best-fit model). When k is uniformly low, the model *under*-estimates both the mean decrease in entropy and the amount of variance around this mean (as indicated by narrower confidence intervals for the model than the experiment). When k is uniformly high, the model *over*-estimates the decrease in entropy for both conditions. These results provide further evidence that the data we observed in the experiment were generated by a population where individuals differ systematically in their sensitivity to the memory load manipulation. Specifically, the superior performance of the u-shaped distribution is suggestive of the nature of these individual differences: in our experiment at least, it seems likely that we were dealing with a population where most people were unaffected by the memory load manipulation, but those who were affected were extremely so.

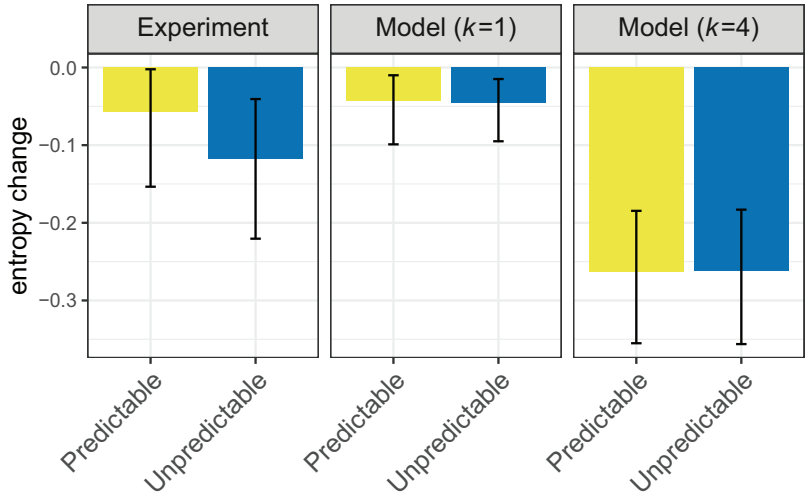

Fig. 8. Change in entropy between input language and production output for participants in the experimental PRODUCTION LOAD conditions and agents in two models with no inter-individual variation in priming strength. These models use the same settings as the best-fit model discussed above for the *priming scope* and *forgetting* parameters; the *population distribution* and *mean priming strength* parameters are not relevant when agents all use the same value of k (the priming strength parameter). When k is low, the model underestimates the true decrease in entropy. When k is high, the model *over*-estimates the decrease in entropy. These models demonstrate the importance of individual differences in priming strength for capturing the experiment results.

Finally, when agents sample from their input with no priming between trials (i.e., $k = 0$ for all agents), there is no significant drop in entropy: results mirror those of the experimental NO LOAD conditions (Fig. 9). This underlines the importance of a production-side mechanism for capturing the experiment results; imperfections in the learning process are not enough to explain the drop in entropy during production.

3.3. Discussion

With this model, we have shown that production mechanisms alone can give rise to levels of regularization comparable to those seen in our experiment, without the need for any prior bias against variability. Specifically, the mechanism implemented by our Pólya urn model can be thought of as a kind of self-priming: rather than agents sampling faithfully from the data they learned, the production process distorts the representation of that data that they draw on during production such that a recently produced variant becomes even more accessible for retrieval in future. Importantly, this distortion is frequency independent: none of our parameter settings involve preferential forgetting of irregular items or preferential retrieval of regular items (cf. Perfors, 2012, where such a model was argued to be the only way that regularization could arise from memory limitations during *learning*). Thus, the skew in the input itself provides the necessary conditions for regularization to occur under a neutral self-priming process.

In our experiment, we saw that the direction of travel was the same for both predictable and unpredictable variation—towards regularity. However, at least descriptively, unpredictable

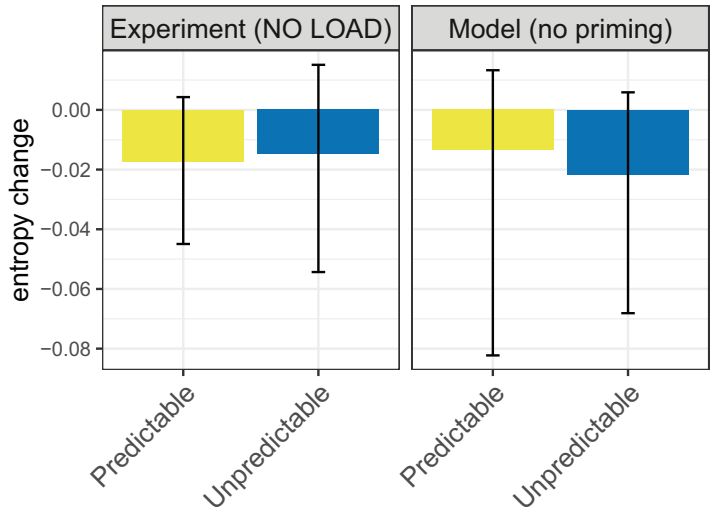

Fig. 9. Change in entropy between input language and production output for participants in the experimental NO LOAD conditions and agents in a model with no priming ($k = 0$ for all agents). When agents sample faithfully from their input, results mirror those of the experimental NO LOAD conditions, that is, no evidence of regularization.

languages tended to change more. One important aspect of these models is that the same parameter settings, applied to the two language types, generate a similar asymmetry. In other words, there is no need to posit different production-side biases targeting the different types of variation: the properties of the input—and differential learning of the two language types—naturally give rise to different amounts of regularization. However, it is true that our models generally perform better in the UNPREDICTABLE condition.

Finally, our results suggest that inter-individual variation in the strength of the priming mechanism is a key ingredient; when priming strength is uniform across the population, the models provide a poor fit to the experiment data. Furthermore, the best model of the population is one in which individuals differ from each other quite radically: most agents fall at one of the two extremes of priming strength, with very few in the middle.

Although our aim here was simply to provide a model that could account for our experimental data, future work could look to apply the mechanism we suggest to other aspects of natural language production that might be relevant to regularization. For example, our experiment does not involve generalization to novel nouns, but this is certainly a task that can increase the tendency to regularization (e.g., Wonnacott & Newport, 2005). Our model could be extended to account for this: the use of a given variant with one noun would prime that variant for all future nouns, whether or not those nouns have been seen before. Furthermore, in both our experiment and model it was not possible to innovate new forms, which removes one potential source of *irregularization*. This could be accounted for in the model through the addition of an error rate parameter which allows for occasional distortions of the sampling process, for example, the addition of an unattested ball to an urn.[17]

4. General discussion

In this study, we have added to a growing body of evidence showing that, even when linguistic variation is accurately learned, it is not always accurately reproduced (e.g., Austin et al., 2022; Ferdinand et al., 2019; Hudson Kam and Chang, 2009; Hudson Kam & Newport, 2009; Saldana et al., 2021; Schwab et al., 2018). Specifically, we have shown that constraints on language production arising from memory limitations can result in the loss of both predictable and unpredictable variation. However, we also found evidence that regularization is not exclusively an effect of production: the process by which random variation becomes conditioned is better explained by learning biases. Humans are powerful statistical learners across many domains, extracting even subtle regularities after very little exposure (see Saffran & Kirkham, 2018, and Sherman, Graves, & Turk-Browne, 2020, for reviews). However, people generally have poor perception of randomness and are quick to infer that random sequences are actually structured (Bar-Hillel & Wagenaar, 1991; Gaissmaier & Schooler, 2008; Hyman & Jenkin, 1956; Wolford, Newman, Miller, & Wig, 2004). Our results suggest that this bias generalizes to language acquisition, causing learners to identify and internalize regularities even when none existed in their input (Samara et al., 2017; Smith & Wonnacott, 2010).

4.1. Memory limitations: Learning or production effects?

In this study, we simulated the memory pressures inherent to language learning and production through a concurrent load task (Hudson Kam, 2019; Perfors, 2012). Of course, language users are not habitually asked to memorize and recall digit sequences during conversation, so this is a somewhat artificial view of working memory's role in language learning and use. Nonetheless, if disrupting working memory during particular linguistic tasks has behavioral consequences, we can infer that memory is a relevant constraint on those tasks generally.

Both our experimental and computational results suggest that memory limitations during language production can account for regularization at the global level (i.e., an overall increase in the frequency of one variant to the exclusion of others) but are not a particularly good predictor of regularization at the lexical level (i.e., the introduction of lexical conditioning). This discrepancy makes sense considering the mechanism that we are proposing for the production effect, whereby variants with a higher frequency (in either the observed data or in the output) become ever more accessible, and therefore ever more likely to be retrieved for production (Goldberg & Ferreira, 2022; Hudson Kam and Chang, 2009; Schwab et al., 2018). Introducing lexical conditioning, on the other hand, requires participants to boost the high-frequency variant for some nouns and the low-frequency variant for others, a process that cannot be easily explained under a memory retrieval account.

In fact, our exploratory analysis suggests that memory limitations during *learning* may have a role to play in explaining the evolution of predictable patterns of variation. At first glance, this result appears to dovetail with some earlier work in the "Less is More" tradition (Newport, 1988; Newport, 1990). For example, simulated agents and recurrent neural networks have been shown to learn linguistic regularities better when they begin with some kind of memory limitation—or input filter—and gradually mature (Elman, 1993; Goldowsky &

Newport, 1993). It has been suggested that, by limiting the size of the sample from which learners can draw inferences, these input filter mechanisms enhance the detection of meaningful relationships (Kareev, 1995; Kareev, Lieberman & Lev, 1997). However, more recent reanalysis (Brooks & Kempe, 2019; Rohde & Plaut, 2003; Rohde & Plaut, 1999) calls many of these findings into question. Specifically, Rohde and Plaut (2003) point out that although filtering mechanisms sometimes isolate the correct regularities, they just as often destroy important parts of the data and identify spurious regularities instead. Indeed, this is exactly what we see here: learners subject to the LEARNING LOAD manipulation are *less* successful at faithfully reproducing the language they were exposed to because they are detecting patterns that did not exist in their input.

Overall, our results lend support to the idea that regularization arises from memory constraints during language production but also suggest that this is not the whole story. If we consider regularization as the process by which language becomes more systematic and predictable—whether by reducing the number of variants in a system, or by specializing different variants for different contexts—then it appears that memory limitations are also doing something important during learning.

4.2. Revisiting the relationship between predictable and unpredictable variation

One of the key aims of this study was to investigate whether linguistic variation is a single phenomenon, with predictable and unpredictable variation constituting two points on the same spectrum. The implication of such a characterization is that the same kinds of biases should act on both types of variation. In other words, the same mechanisms that have been shown to result in regularization of unpredictable variation should also target predictable variation. Our results are consistent with this account when it comes to the effect of memory limitations during language *production*.

However, our analysis also suggests that truly random variation is subject to distortion during the learning process in a way that conditioned variation is not—even when that conditioning is only probabilistic. In other words, even though learning biases *could* theoretically have obscured the small amount of noise in our predictable languages, in fact participants' estimates show that they were very aware of this noise and did not believe that they had been exposed to a deterministic pattern of conditioning. Therefore, it appears that there may be something special about unpredictable variation when it comes to learning. Specifically, our results suggest that language learning is biased in favor of predictable dependencies between elements in a system to the extent that even random systems will be analyzed as containing such patterns. Future research could investigate how these learning biases play out across the spectrum of variation; for example, a less deterministic version of our predictable language might be subject to more distortion in learning.[18]

We observed two related but distinct biases in this study: a bias against variability of all kinds (driven by production) and a bias against unpredictability (driven by both learning and production). However, the second of these biases appears to be stronger: In both learning and production, we saw much bigger changes in MI than in entropy. A question for future work is how the relative strength of these biases interacts with the size of the system: with only two

variants, as in our design, it is presumably not difficult to maintain both. Expanding the language may heighten the pressure to regularize by losing some variants altogether, rather than just by introducing conditioning (although see Hudson Kam & Newport, 2009). Our forced-choice production task also minimized the kind of retrieval difficulties that we might expect to result in increased use of one variant to the exclusion of others since participants were cued to remember that there was another option even if they would have spontaneously favored a single variant. We would expect a different kind of production task—with participants required to free-type or orally produce their descriptions—to give rise both to a greater drop in entropy overall (Hudson Kam and Chang, 2009), and to a stronger effect of the interference task, especially if the language was more complex.

Overall, our results suggest that pragmatic factors alone cannot fully explain regularization and that working memory limitations offer a plausible *cognitive* explanation for this phenomenon. Specifically, we found that increased cognitive load during language production gave rise to increased regularization of both predictable and unpredictable variation—in the absence of any communication between participants or differences in pragmatic framing of the task. Furthermore, a pragmatic account would not predict any regularization during learning, since the mechanisms implicated in such accounts only come into play during production. However, our results clearly show that when participants produce a more predictable language than the one they were exposed to, this is at least partly because they have failed to accurately learn the randomness in their input.

4.3. Why do languages have variation at all?

Our results suggest that biases arising from memory limitations broadly disfavor linguistic variation, even when that variation is predictable. From the perspective of language evolution, one might, therefore, wonder why variation is so pervasive in natural languages. As with any cognitive bias shaping language, the explanation for this is likely a combination of the fact that these biases are weak (i.e., defeasible) and compete with other pressures shaping language. Most obviously, patterns of linguistic usage are influenced by the social contexts in which they are found: there is ample evidence to suggest that speakers use variation as a marker of social identity (see Chambers & Schilling, 2018, for an overview). Furthermore, some types of variation may be preferred because of cognitive biases pertaining to specific linguistic or semantic categories (e.g., Christensen, Fusaroli, & Tylén, 2016; Holtz, Kirby, & Culbertson, 2022; Motamedi, Wolters, Naegeli, Kirby, & Schouwstra, 2022; Napoli & Sutton-Spence, 2014; Schouwstra & De Swart, 2014).

Individual differences in the strength of the regularization bias may also help to explain how variation can persist in natural language. Our experimental data certainly suggest that memory load does not lead to regularization across the board. In particular, a wide range of behaviors were represented in our PRODUCTION LOAD conditions. Many participants in these conditions seemed not to be hindered at all by the interference task, producing languages with near-zero entropy change compared to the input.[19] Some appeared to be moderately affected, maintaining some but not all of the variation that was present in the input. And a small handful were severely disrupted, producing only one variant in testing. Similarly, our

best-fit computational model was one in which the majority of agents actually had no propensity towards regularization, but those who did tended to reduce variation quite substantially. In terms of diachronic change in natural language then, if only some individuals have very strong biases against variation, we should perhaps not expect that variation to be lost either quickly or completely.

Finally, although this was not relevant in our study, systems of conditioned variation may persist due to frequency-dependent patterns in regularity. Irregular forms tend to be highly frequent, presumably making it easier to learn and retrieve the correct form (Cuskley et al., 2014; Wu, Cotterell, & O'Donnell, 2019). Furthermore, learners are sensitive to the frequency of specific exemplars (e.g., the frequency of the word *went*) as well as the frequency of morphological types (e.g., the frequency of the *-ed* past tense marker), so it is not necessarily the case that the "regular" variant is the most easily retrieved in all contexts (Arnon, 2015; Arnon & Snider, 2010). Indeed, usage-based models (e.g., Bybee, 2006; Bybee, 2002; Hay, 2001; Langacker, 1988) argue that the easiest variant to access in any given context is simply the one that has been experienced most often in that context. In such models, linguistic data form memory representations whereby items that are experienced frequently together start to form a unit; these units then come to be processed and retrieved holistically and thus become resistant to restructuring (Bybee, 1985; Bybee & Thompson, 1997). There is also growing recognition that learners actually *start out* with such holistic units in some cases, especially for high-frequency items (Arnon & Clark, 2011; Chevrot, Dugua, & Fayol, 2008; Christiansen & Arnon, 2017; Havron & Arnon, 2021; Lieven, Pine, & Baldwin, 1997; Pine & Lieven, 1997; Siegelman & Arnon, 2015). In this case, lexically conditioned variation may persist because highly frequent irregular items never get segmented, and thus, when producing these items, there is no process of retrieving individual morphemes during which an alternative form could be retrieved (cf. Pinker & Ullman, 2002). Therefore, while regularization might arise when a high-frequency type is extended to a less familiar context (Harmon & Kapatsinski, 2017; Koranda, Zettersten, & MacDonald, 2018; Wonnacott, 2011), high-frequency irregular items are likely to be evolutionarily stable. All nouns in our design were equally frequent, so our results do not speak to any potential relationship between frequency and memory limitations in driving regularization. However, the paradigm we present here could certainly be used to test this hypothesis.

5. Conclusion

We have provided evidence that cognitive biases leading to regularization target both unpredictable and predictable variation. Our findings support the idea that regularization is particularly strong during production and is driven at least in part by memory limitations. However, our results also suggest that this is not the whole story; while over-retrieval of a more accessible variant during language production may act to reduce overall variability, *unpredictability* appears to decrease more as a result of inferences formed during learning. Overall, this study lends support to the notion that cognitive constraints in individuals can give rise to particular structures in languages. Specifically, we argue that—all things

equal—regularities that allow languages to pass more easily through the bottleneck imposed by working memory limitations will tend to accumulate as languages evolve, leading to the appearance of typological universals.

Acknowledgments

This project was supported by funding from the Economic and Social Research Council (grant ref. ES/P000681/1, awarded to AK) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643, awarded to JC). We are grateful to Elizabeth Pankratz for drawing the stimuli, and to members of the Centre for Language Evolution for helpful discussion. Thank you to our reviewers, Carla Hudson Kam and two anonymous reviewers, for valuable input.

Open Research Badges


This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/9e27b

Notes

1 Either by producing an invalid label type (*noun* + *noun*, *marker* + *marker*, or *marker* + *noun*; the only valid label type was *noun* + *marker*), or by producing an incorrect noun.
2 Defined as clicking buttons in the same left-right position on more than 90% of trials.
3 Due to a technical error, the order of presentation was not fully randomized in this phase. Instead, all participants saw eight passes through the stimuli set in the same randomized order each time. We have no reason to expect that this would have affected participant behavior.
4 This potentially reduces the strength of regularization behavior compared to a free production task, but it is still a task where regularization can be observed with the right analysis techniques, for example, Ferdinand et al. (2019)
5 For example, due to primacy or recency effects (Ferdinand et al., 2019).
6 $H(V) - H(V|C)$.
7 Note that, although participants could in principle produce a language with MI of 1, this is not what we would see if they simply produced a deterministic version of the conditioning pattern in their input (i.e., four nouns with one marker and two with the other); such a language would score 0.92.
8 The pattern of results under this analysis is identical to the one obtained from our preregistered linear models.
9 Note that we do not draw any inferences from significantly positive z-scores for entropy change or significantly negative z-scores for MI change: none of our predictions are

about an increase in variability, so our focus is simply on whether there is or is not evidence for regularization. In other words, these are all one-tailed tests.

10 Shuffling in this way will, on average, give the same mean in each condition, so the resulting distribution will be normal and centered around 0, that is, no difference between conditions.

11 This resulted in the exclusion of 322 (out of 8,433) trials. Of these, the word occupying the noun slot was an incorrect noun on 267 trials, of which the label was a valid *noun + plural* form on 168 trials; on the remaining 99 trials, the word occupying the plural slot was another noun, suggesting that the participant had tried to correct their mistake with their second click (as indicated in some debrief questionnaires). Of the remaining 55 trials, there were 34 cases where the noun was correct but the label was invalid because the noun had been duplicated. In 11 cases, the noun was correct but the words were in the wrong order (i.e., the label was of the form *plural + noun*).

12 Note that there was more scope for MI to increase in UNPREDICTABLE conditions because the starting point (0) was lower for these languages than in PREDICTABLE conditions (0.41).

13 In this case, we compare the real by-condition means to the mean of a corresponding simulated condition where participants probability match their *estimates* (rather than the input).

14 Rounding up for "ruffo": this slider was set to 99%.

15 A mean greater than 4 would allow k to take values outside of the defined range.

16 Strictly speaking, this parameter controls the *median* of the u-shaped distribution rather than the mean, since the distribution is asymmetric.

17 Thank you to an anonymous reviewer for these interesting suggestions.

18 Thank you to an anonymous reviewer for this suggestion.

19 Although we would expect to see fewer participants in this category if the production task itself was more taxing, that is, free production rather than forced-choice.

References

Aitchison, J. (1996, July 20). Small steps or large leaps? Undergeneralization and overgeneralization in Creole acquisition. In H. Wekker (Ed.), *Creole languages and language acquisition* (pp. 9–32). Berlin, Germany: Mouton De Gruyter.

Arnon, I. (2015). What can frequency effects tell us about the building blocks and mechanisms of language learning? *Journal of Child Language*, *42*(2), 274–277.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth - children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*(2), 107–129.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Austin, A. C., Schuler, K. D., Furlong, S., & Newport, E. L. (2022). Learning a language from inconsistent input: Regularization in child and adult learners. *Language Learning and Development*, *18*(3), 249–277.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation - Advances in Research and Theory*, *8*(100), 47–89.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, *12*(4), 428–454.

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

Bickerton, D. (1981). *Roots of language*. Berlin, Germany: Language Science Press.

Braine, M. D. S., Brody, R. E., Brooks, P. J., Sudhalter, V., Ross, J. A., Catalano, L., & Fisch, S. M. (1990). Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language*, *29*(5), 591–610.

Brooks, P. J., & Kempe, V. (2019). More is more in language learning: Reconsidering the less-is-more hypothesis. *Language Learning*, *69*, 13–41.

Brown, H., Smith, K., Samara, A., & Wonnacott, E. (2022). Semantic cues in language learning: An artificial language study with adult and child learners. *Language, Cognition and Neuroscience*, *37*(4), 509–531.

Bybee, J. (1985). *Morphology: a study of the relation between meaning and form*. Amsterdam, The Netherlands: John Benjamins.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*, 261–290.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. *Annual Meeting of the Berkeley Linguistics Society*, *23*(1), 378–388.

Carroll, R., Svare, R., & Salmons, J. C. (2013). Quantifying the evolutionary dynamics of German verbs. *Journal of Historical Linguistics*, *2*(2), 153–172.

Chambers, J. K., & Schilling, N. (2018). *The handbook of language variation and change*. Hoboken, NJ: John Wiley & Sons.

Chevrot, J.-P., Dugua, C., & Fayol, M. (2008). Liaison acquisition, word segmentation and construction in French: A usage-based account. *Journal of Child Language*, *36*(3), 557–596.

Christensen, P., Fusaroli, R., & Tylén, K. (2016). Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization. *Cognition*, *146*, 67–80.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*(3), 542–551.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*(2), 317–335.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2019). Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias. *Language*, *95*(2), 268–293.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*, 1964.

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82.

Culbertson, J., Smolensky, P., & Wilson, C. (2013). Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, *5*(3), 392–424.

Culbertson, J., & Wilson, C. (2013). Artificial grammar learning of shape-based noun classification. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 2118–2123.

Cuskley, C., Castellano, C., Colaiori, F., Loreto, V., Pugliese, M., & Tria, F. (2017). The regularity game: Investigating linguistic rule dynamics in a population of interacting agents. *Cognition*, *159*, 25–32.

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., & Tria, F. (2014). Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English. *PLoS ONE*, *9*(8), e102882.

DeGraff, M. (1999). *Language creation and language change: Creolization, diachrony, and development*. Cambridge, MA: MIT Press.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, *83*, 37–64.

Fehér, O., Ritt, N., & Smith, K. (2019). Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language*, *109*, 104036.

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.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218–245.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, *90*(1), 58–89.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422.

Gardner, R. A. (1957). Probability-learning with two and three choices. *The American Journal of Psychology*, *70*(2), 174–185.

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

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four… or is it two? *Memory*, *12*(6), 732–747.

Goldberg, A., & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, *26*(4), 300–311.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The less is more hypothesis. *The Proceedings of the 24th Annual Child Language Research Forum* (pp. 124–138). Stanford, CA: Stanford Linguistics Association by the Center for the Study of Language and Information.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology*, *98*, 22–44.

Havron, N., & Arnon, I. (2021). Starting big: The effect of unit size on language learning in children and adults. *Journal of Child Language*, *48*(2), 244–260.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*(6), 1041–1070.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Holtz, A., Kirby, S., & Culbertson, J. (2022). The influence of category-specific and system-wide preferences on cross-linguistic word order patterns. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. *44*, pp. 1011–1018). Austin, TX: Cognitive Science Society.

Hudson Kam, C. L. (2015). The impact of conditioning variables on the acquisition of variation in adult and child learners. *Language*, *91*, 906–937.

Hudson Kam, C. L. (2019). Reconsidering retrieval effects on adult regularization of inconsistent variation in language. *Language Learning and Development*, *15*(4), 317–337.

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(3), 815–821.

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

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

Hyman, R., & Jenkin, Noel. S. (1956). Involvement and set as determinants of behavioral stereotypy. *Psychological Reports*, *2*(3), 131–146.

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*(3), 335–352.

Kamps, C., Ferdinand, V., & Kirby, S. (2014). The origins of regularity in language: Why coordination matters. In Cartmill, E. A., Roberts, S., Lyn, H., & Cornish, H., (Eds.), *The evolution of language: Proceedings of the 10th International Conference* (pp. 457–458). Singapore: World Scientific

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*(3), 263–269.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General*, *126*, 278–287.

Karmiloff-Smith, A. (1981). *A functional approach to child language: a study of determiners and reference*. Cambridge, England: Cambridge University Press.

Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford, England: Oxford University Press.

Koranda, M., Zettersten, M., & MacDonald, M. C. (2018). Word frequency can affect what you choose to say. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 40, pp. 629–634). Madison, WI: Cognitive Science Society.

Langacker, R. W. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics* (pp. 127). Amsterdam. The Netherlands: John Benjamins

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713–716.

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*(1), 187–219.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226.

Mahmoud, H. (2008). *Polya urn models*. Boca Raton, FL: CRC Press.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), 1–178.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Motamedi, Y., Wolters, L., Naegeli, D., Kirby, S., & Schouwstra, M. (2022). From improvisation to learning: How naturalness and systematicity shape language evolution. *Cognition*, *228*, 105206.

Napoli, D. J., & Sutton-Spence, R. (2014). Order of the major constituents in sign languages: Implications for all language. *Frontiers in Psychology*, *5*, 376.

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American sign language. *Language Sciences*, *10*(1), 147–172.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, *67*(4), 486–506.

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development*, *12*(2), 138–155.

Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*(2), 123–138.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463.

Pérez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. *Journal of Child Language*, *18*(3), 571–590.

R Core Team. (2022). R: A language and environment for statistical computing. Vienna, Austria: R Project for Statistical Computing.

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Sciences*, *11*(7), 274–279.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*(6), 927–960.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Rische, J. L., & Komarova, N. L. (2016). Regularization of languages by adults and children: A mathematical framework. *Cognitive Psychology*, *84*, 1–30.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67–109.

Rohde, D. L. T., & Plaut, D. C. (2003). Less is less inlanguage acquisition. In P. Quinlin (Ed.), *Connectionist modelling of cognitive development* (pp. 160–200). East Sussex, England: Psychology Press.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203.

Saldana, C., Smith, K., Kirby, S., & Culbertson, J. (2021). Is regularisation uniform across linguistic levels? Comparing learning and production of unconditioned probabilistic variation in morphology and word order. *Language Learning and Development*, *17*(2), 158–188.

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

Schouwstra, M., & De Swart, H. (2014). The semantic origins of word order. *Cognition*, *131*(3), 431–436.

Schwab, J. F., Lew-Williams, C., & Goldberg, A. (2018). When regularization gets it wrong: Children over-simplify language input only in production. *Journal of Child Language*, *45*(5), 1054–1072.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological Science*, *12*(4), 323–328.

Senghas, A., Coppola, M., Newport, E. L., & Supalla, T. (1997). Argument structure in Nicaraguan sign language: The emergence of grammatical devices. In Hughes, E., Hughes, M., & Greenhill, A., (Eds.), *Proceedings of the Boston University Conference on Language Development* (Vol. *21*, pp. 550–561)

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences*, *32*, 15–20.

Siegel, J. (2007). Recent evidence against the language bioprogram hypothesis. *Studies in Language*, *31*(1), 51–88.

Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60–75.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American sign language from inconsistent input. *Cognitive Psychology*, *49*(4), 370–407.

Smith, K., Ashton, C., & Sims-Williams, H. (2023). The relationship between frequency and irregularity in the evolution of linguistic structure: An experimental study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*, 851–857.

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: Biological Sciences*, *372*, 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.
Smith, K. H. (1969). Learning Co-occurrence restrictions: Rule induction or rote learning? *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 319–321.
Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, *41*(3), 623–658.
Walsh, M., Möbius, B., Wade, T., & Schütze, H. (2010). Multilevel exemplar theory. *Cognitive Science*, *34*(4), 537–582.
Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology*, *58*(4), 221.
Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*(1), 1–14.
Wonnacott, E., & Newport, E. L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In Brugos, A., Clark-Cotton, M., & Ha, S. (Eds.), *BUCLD 29: Proceedings of the 29th Annual Boston University Conference on Language Development*, (pp. 1–11). Somerville, MA: Cascadilla Press.
Wu, S., Cotterell, R., & O'Donnell, T. (2019). Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5117–5126). Stroudsburg, PA: Association for Computational Linguistics.

Appendix A: Individual-level experimental data

All plots in this appendix show individual participants as colored points and condition means as black points. Error bars represented bootstrapped 95% confidence intervals over the mean.

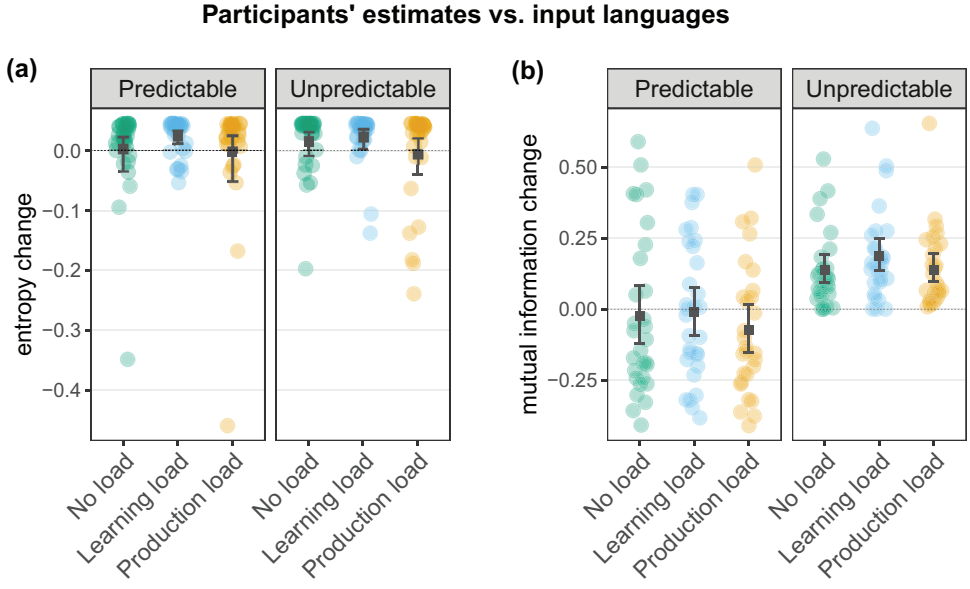

Fig. A.1. Change in entropy (left) and MI (right) between the languages participants were trained on and the ones described by their estimates, by condition.

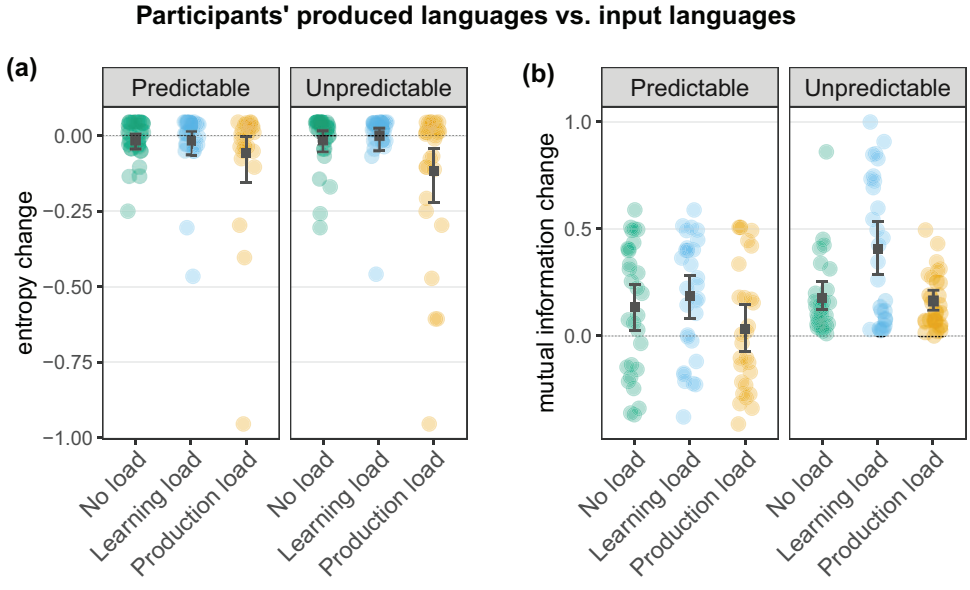

Fig. A.2. Change in entropy (left) and MI (right) between the languages participants were trained on and the ones they produced, by condition.

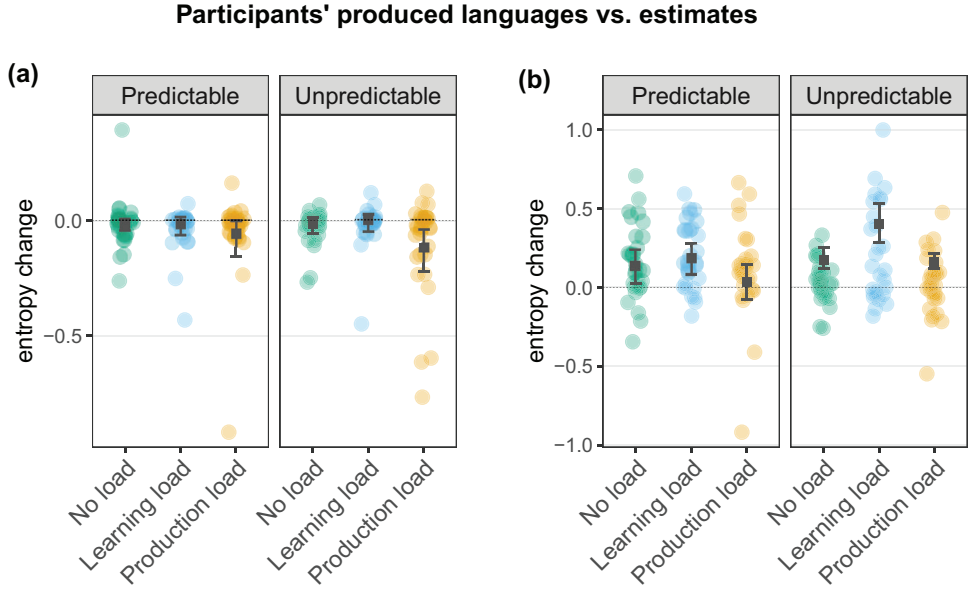

Fig. A.3. Change in entropy (left) and MI (right) between the languages described by participants' estimates and the ones they produced, by condition.

Appendix B: Additional model analysis

In general, all computational models were closer to the real data on entropy than MI, meaning that they were capturing the overall loss of variation better than the increase in lexical conditioning. Different settings of the *priming scope* parameter in particular generated very different results for the two measures, as shown in Table B.1. On average, "within-nouns" models performed considerably better than others on entropy. However, these models dramatically overestimated the change in MI relative to the experiment, since priming only within nouns leads to a very high likelihood of lexical conditioning (i.e., an increase in MI). In fact, the single best-fit model to the entropy data (divergence = 0.046) provided one of the *worst* fits to the MI data (divergence = 1.103), meaning that it was impossible to select a single model that could capture both effects in the experiment. "Between-nouns" models exhibited the opposite problem: while some models provided a reasonable fit for the UNPREDICTABLE condition, MI always *decreased* more in the PREDICTABLE condition than in the real experiment. Although "between-nouns" models had the lowest average divergence score overall, the single best-fitting model used the "within-and-between" setting. Moreover, these models had the most similar performance between entropy and MI.

The performance of different settings for the *mean priming strength* and *forgetting* parameters depended heavily on *priming scope* and varied between measures. For entropy, there was a negative correlation between *mean priming strength* and average divergence scores for "within-nouns" models and a positive correlation for others. In other words, higher means provided a better fit for "within-nouns" models, while lower means performed better for "between-nouns" and "within and between" models. Similarly, models with "oldest" *forgetting* performed marginally better than "random" models when priming was only within nouns, but considerably worse when priming was between nouns or both within and between. Overall, averaging over different settings of the *priming scope* parameter, higher means (Table B.2), and "oldest" forgetting (Table B.3) always provided a worse fit.

In terms of the *population distribution* parameter, there was relatively little difference between uniform and normal-like models (especially on entropy), but u-shaped models considerably out-performed both across the board (Table B.4). In fact, the top 10 best-fitting models overall all used the u-shaped distribution.

Table B.1
Divergence scores for different settings of the *priming scope* parameter

Priming Scope	Entropy	MI	Overall
Within nouns	0.122	0.618	0.370
Between nouns	0.298	0.374	0.336
Within and between	0.411	0.334	0.372

Abbreviation: MI, mutual information.

Table B.2
Divergence scores for different settings of the *mean priming strength* parameter

Mean priming strength	Entropy	MI	Overall
1.0	0.130	0.310	0.220
1.5	0.156	0.361	0.258
2.0	0.205	0.408	0.307
2.5	0.264	0.447	0.356
3.0	0.325	0.488	0.407
3.5	0.395	0.524	0.459
4.0	0.462	0.555	0.508

Abbreviation: MI, mutual information.

Table B.3
Divergence scores for different settings of the *forgetting* parameter

Forgetting	Entropy	MI	Overall
Random	0.149	0.376	0.262
Oldest	0.405	0.508	0.456

Abbreviation: MI, mutual information.

Table B.4
Divergence scores for different settings of the *population distribution* parameter

Population distribution	Entropy	MI	Overall
Uniform	0.306	0.476	0.391
Normal-like	0.306	0.537	0.421
U-shaped	0.219	0.313	0.266

Abbreviation: MI, mutual information.