Cognition 212 (2021) 104512

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Familiarity, consistency, and systematizing in morphology

R. Alexander Schumacher [a,*], Janet B Pierrehumbert [a,b]

[a] Northwestern University, USA
[b] University of Oxford, UK

ARTICLE INFO

Keywords:
Inconsistency
Language change
Artificial language learning
Regularization
Inter-participant variability
Morphology

ABSTRACT

Language learning involves exposure to inconsistent systems - that is, systems where multiple patterns or methods exist to mark some meaning. Inconsistent systems often change to be more regular over time - they become systematized. However, some recent studies have reported that learners tend to reproduce inconsistency in the input, leading to models in which the language learning mechanism is basically preservatory. We ran an artificial language learning experiment using a novel paradigm to extend our understanding of systematizing versus preservatory mechanisms in language learning. Participants were taught two number marking systems, either completely consistently (the probability P of the system is 1.00) or inconsistently (with $P = 0.875$ for one system and $P = 0.125$ for the other, and so on for $P = 0.75$ and $P = 0.625$). One marking system was a plural-marking system. The other was a typologically rare singulative-marking system. When generalizing to novel items, participants produced more regular output patterns overall for more consistent input conditions than for less consistent ones, and more for the plural-marking conditions than for the singulative-marking conditions. For the singulative-marking conditions, the inter-participant variation was much greater than for the plural-marking ones; some individuals systematized towards the more familiar pattern, some systematized towards the less familiar pattern and some were not significantly different from probability-matching. We analyze the variation in relation to current statistical learning models, showing that preservatory learning models, as well as all models with a single free parameter, fail to capture our results. We show how a model with two free parameters in which individuals can vary in their propensity to systematize in any given situation is more successful. We also discuss implications for the theory of language change.

1. Introduction

Natural languages are highly structured, but they are never perfectly regular. Language learners may encounter conflicting patterns in their experience for many reasons. Patterns may be variable across speakers, because of dialectal differences or non-native speakers within the community, for example. Within-speaker patterns often vary during periods of language change. Systemic inconsistencies arise when some words in a language follow one generalization, and other words follow a different and conflicting generalization. These have been extensively studied in linguistics under the rubric of rules, minor rules, and exceptions in the domains of syntax, morphology, and morphophonology (Jackendoff, 1975; Lakoff, 1971). When multiple, different, systems exist in a learner's experience, the situation is one in which the learner has encountered *inconsistency*.

How learners process and encode inconsistency is a major issue in psycholinguistics, with ramifications for the theory of language variation and change and ultimately linguistic typology. For example, inconsistency is ubiquitous during early stages of the development of creoles from pidgins. At historic time scales, poorly structured pidgins that arise through initial language contact situations tend to evolve towards more structured creoles (Hudson Kam & Newport, 2009; Thomason & Kaufman, 1988). Language changes such as changes in basic word order tend to follow an S-shaped pattern to complete regularity, rather than stabilizing at some intermediate level of adoption (Blythe & Croft, 2012; Kroch, 1989; Labov, 1994). Because such developments arise from the learning and use of language by individuals in a speech community, these broad trends likely reflect cognitive tendencies towards regularity. The observation that some inconsistent systems become regular very slowly and that some variation persists for many generations, however, raises questions about when and to what extent tendencies towards regularity come into play.

Recently, artificial language learning experiments have emerged as a powerful paradigm for exploring how language learners process and encode inconsistency (Culbertson, Smolensky, & Legendre, 2012; Ferdinand, Kirby, & Smith, 2019; Hudson Kam & Newport, 2005; Kirby,

* Corresponding author.

https://doi.org/10.1016/j.cognition.2020.104512
Received 3 March 2017; Received in revised form 11 October 2020; Accepted 5 November 2020

Cornish, & Smith, 2008; Reali & Griffiths, 2009; Smith & Wonnacott, 2010; van de Vijver & Baer-Henney, 2014; Vouloumanos, 2008; Wonnacott & Newport, 2005). Some tendency towards regularity was observed in each study. However, only a few studies (such as Smith & Wonnacott, 2010, Fedzechkina, Jaeger, & Newport, 2012, and Schuler, Yang, & Newport, 2016) use artificial language learning to examine one of the most interesting areas for the investigation of the learning of inconsistency, namely inconsistency in morphology.

Inconsistency at the morphological level is interesting and important because it is a source of inconsistency that language learners often encounter, and much of this inconsistency is systemic. For example, a regular pattern of affixation of *-ed* to English verbs to form the past tense co-exists with several vowel alternation patterns that also express the past tense, as in *break/broke, freeze/froze*, and so on. While these minority patterns (which represent the opaque residue of a Germanic vowel alternation) only characterize a small number of verbs, they can nonetheless generalize to novel verbs that are highly similar, indicating that learners have formed a minor generalization that conflicts with the primary one (Albright & Hayes, 2003; Rácz, Beckner, Hay, & Pierrehumbert, in press). Because morphology frequently presents learners with inconsistency, it is a novel and significant angle from which to investigate how learners process and encode inconsistent systems. In particular, it is possible that learners may have different expectations concerning inconsistency in a domain of their experience like morphology where it is comparatively common and known to be cognitively encoded. Irregularity in morphology not only represents an area where adults know an inconsistent system, but where such systems are a common consequence of the natural evolution of a language.

Though an extensive literature explores learning of inconsistent morphological systems by children (c.f. Bybee & Slobin, 1982; Marcus et al., 1992; Plunkett & Marchman, 1992), the potential of artificial language learning for exploring morphological learning by adults has not yet been fully realized. As pointed out by Bybee and Beckner (2010), adults participate in and contribute to language change (see Kaschak & Glenberg, 2004; Sankoff & Blondeau, 2007; Wagner & Sankoff, 2011). In our study, we look at the effects of inconsistency in the adult learning of number marking morphology. We compare two number-marking systems. One is the typologically common plural-marking pattern that is highly familiar to the participants because it is the pattern used in English. The other is a pattern that is typologically rare and not used at all in English (the singulative-marking system to be described below in Section 3.2). It is accordingly less familiar to adult English-speaking participants (if indeed it is familiar at all). Our study used a novel paradigm that gathers information about the time course of learning, which allows us to compare how adults learn more familiar or less familiar systems that present with variation. With the novel manipulation and training paradigm, we have three main goals. First, we seek to better understand the conditions under which learners will reduce or preserve inconsistency in the input. Second, we seek to understand whether or not learners are biased or unbiased in whether they will preserve or reduce inconsistency, and finally, we seek to characterize the inter-participant variability observed in this learning. As we will see, the results differ substantially from those of previous artificial language learning experiments that evaluate the effects of inconsistency. The differences may be due to differences in the domain (morphology as opposed to syntax, phonology, or word learning), or they may be due to differences in the experimental paradigm; we return to this issue in the discussion.

2. Background and motivation

2.1. Output classification in artificial language learning

The way that learners process and encode inconsistency can be assessed by examining the relationship of their output to the input they have experienced. The results of many studies have attached significance to two common output patterns that learners may exhibit: *probability matching* (also called *frequency tracking*) and *regularization* (Culbertson et al., 2012; Hudson Kam & Newport, 2005, 2009; Kirby, 2001; Kirby et al., 2008; Perfors, 2012; Reali & Griffiths, 2009; Smith & Wonnacott, 2010; van de Vijver & Baer-Henney, 2014; Vouloumanos, 2008; Wonnacott & Newport, 2005).

Probability matching occurs when learners reproduce the input frequencies in their output. For example, if learners were presented with a language in which the past tense is marked by *-ot* with probability 0.7 and *-oo* with probability 0.3, a learner who produces probability matching would produce *-ot* at about $P = 0.7$ and *-oo* at about $P = 0.3$. A significant number of studies have reported probability matching. One example is Hudson Kam and Newport (2005), who manipulated the presence/absence of determiner forms in an artificial language with a phrasal syntax. Determiners were presented with nouns at rates of 0.45, 0.60, 0.75, and 1.00. On the average, adult participants matched the input frequencies of determiner occurrence to approximately the same extent in each condition. Vouloumanos (2008) found that learners match probabilities for inconsistency in object labelling. Learners were shown six different objects ten times, seeing one of the two labels associated with the object, with rates of 0.50, 0.60, 0.70, 0.80, 0.90, and 1.00. Learner's behavior did not differ significantly from probability matching. Smith and Wonnacott (2010) found similar results for inconsistency in the use of a number-marking affix.

Learners may also regularize the input. Regularizing reduces the amount of variation by favoring one variant over others (Hudson Kam & Newport, 2005). In the example above, a person who produces *-ot* significantly more than 0.70 of the time, with *-oo* occurring correspondingly less often, would be characterized as a regularizer. Assuming that the learner's goal is to correctly predict the form of unseen expressions, the optimal choice is to produce only the most frequent variant (*majority regularization*), since it is expected to succeed more often than reproducing the input frequencies would. However, regularization might also be achieved by favoring the lower frequency variant so much that it is the majority variant in the output, and the output has less variation than input (*minority regularization*). So, a minority regularizer would output *–oo* more than 0.70 of the time, and *–ot* the rest of the time. This outcome is observed for some children in Hudson Kam and Newport (2009), and for some adults in the Baer-Henney, Kügler, and van de Vijver (2014) study of vowel co-occurrence rules.

Majority regularization has been observed in some studies.[1] Wonnacott and Newport (2005) found regularization when participants generalized input probabilities to novel vocabulary. Participants learned nouns and verbs, and were then exposed to the syntax of an artificial language with two dominant word orders (two thirds of the examples had the VSO order, and one third had the VOS order). Learners regularized the word order to VOS or VSO for vocabulary that they had not trained on, while they produced probability matching on the vocabulary that they had been trained on. Hudson Kam and Newport (2009) examined the behavior of participants as the number of competing forms increased. Manipulating the word forms of the determiners, they presented participants with a dominant determiner (occurring at a rate of 0.60) and then filled the remaining exposures with "noise" determiners of other frequencies (2, 4, 8, or 16 distinct competitors). Regularization of the dominant determiner increased as the number of competing forms increased. Another significant study where regularization was found is Culbertson et al. (2012). They taught participants different orders of modifiers with respect to a noun. Modifiers were ordered before or after the noun, and the experimental conditions varied according to which of these orders for either adjective or numeral modifiers was dominant ($P = 0.70$). They hypothesized that since one order (Adj-Noun-Num) is typologically rare, participants would be biased against it. Participants

[1] Ferdinand (2015) has a comprehensive review of studies which describe regularization.

regularized typologically well-attested orders, while they did not regularize the typologically rare word order.

A fourth possible pattern is *irregularization*. We define irregularization as occurring when a more frequent variant is dispreferred in the output, but not so much that the output has less variation (*a la* minority regularization). Consequently, the output has more variation than the input. Below is a table (Table 1) illustrating the possible outcomes, using the *-ot/-oo* example.

2.2. Mechanisms of language learning

The exact nature of the mechanism that determines what output pattern a learner will produce is controversial. On one view, the learning mechanism is fundamentally *preservatory;* when learners encounter inconsistency, they encode it as inconsistent, tending to preserve in their output the frequencies of the different variants that they experienced. The outcomes that deviate from probability matching (majority regularization, minority regularization and irregularization) are outcomes that implicate additional factors.

Two such factors have been identified in the literature. One is memory limitations. Hudson Kam and Newport (2009) and Hudson Kam and Chang (2009) suggest that very low levels of exposure to a variant can lead to incomplete learning, with the result that the variant is not available for later use. This mechanism can increase the probabilities of the more common variants. However, a similar effect comes about in models with multiple generations of learners simply through sparse sampling effects. In work on language evolution, the baseline neutral evolution model (in which no variants or interlocutors are favored) presupposes a preservatory learning mechanism. The input to each generation is a finite random sample generated using the probability estimates acquired by the previous generation. This model has been extensively analyzed (Baxter, Blythe, Croft, & McKane, 2009; Blythe & Croft, 2012; Ferdinand, 2015; Ferdinand et al., 2019). If a variant has a run of bad luck during production, and fails to occur in the input to the next generation, it will be lost. Over multiple generations, the asymptotic behavior is regularization to a single variant (referred to as "fixation" of that variant), and the probability that a variant will be fixated is proportional to its initial frequency. Thus, neutral evolution favors majority regularization, but it can generate minority regularization – albeit with extremely low probability. The multi-generational model in Reali and Griffiths (2009) also relies on this mechanism, with the additional factor of a Bayesian prior that influences each generation anew.

Another factor that might cause deviation from probability matching is *substantive bias.* A *substantive bias* favors variants that have some characteristics over others that lack these characteristics. The bias could arise from the previous linguistic experience of the learner (Baer-Henney et al., 2014; Janse & Newman, 2013). It could arise from functional factors, such as informational efficiency, phonetic naturalness, or phonetic robustness (Hayes, Siptar, Zuraw, & Londe, 2009; Niyogi, 2006; van de Vijver & Baer-Henney, 2012). Substantive biases could also arise from social factors, such as a propensity to produce variants that are associated with admired or well-connected people (Blythe & Croft, 2012; Fagyal, Swarup, Escobar, Gasser, & Lakkaraju, 2010). A further possibility, associated with nativist theories of language, is that Universal Grammar may exercise an influence during adult language learning (Culbertson et al., 2012; Culbertson & Smolensky, 2012). Findings that learners are biased towards typologically common patterns, even when these patterns are unattested in their own language, are interpreted as favoring an operative role for Universal Grammar. Whatever the source of a substantive bias, its effect on a preservatory learning model will be to increase the frequency of any variant with the key characteristics, regardless of its initial frequency. This can result in majority regularization, if the variant was already dominant. A strong bias can also result in irregularization or minority regularization, if the favored variant is not the dominant one.

Table 1
Sample outcomes for each of the four output patterns discussed. Input pattern: P(−ot) = 0.70 and P(−oo) = 0.30.

Output Classification	Productions of *-ot*	Productions of *-oo*
Probability matching	~0.70	~0.30
Regularization (majority favored)	>0.70	<0.30
Regularization (minority favored)	<0.30	>0.70
Irregularization	between 0.30 and 0.70	between 0.30 and 0.70

An alternative view about regularization is that it is an intrinsic property of language learning – learners have a penchant for *systematicity*. As they encounter language input, they are implicitly constructing abstract grammatical rules or constraints that prune out or reduce inconsistencies. For a systematizing mechanism, inconsistent systems are unstable, and so regularization is the more expected output pattern. Systematizing can be captured through use of a nonlinear function directly relating the input frequencies to the output frequencies. Specific proposals for the form of such a function include the sigmoidal functions developed in Ashby and Maddox (1993), Kirby, Dowman, and Griffiths (2007), Mandelshtam and Komarova (2014), and Pierrehumbert, Stonedahl, and Daland (2014). Similar effects are achieved in models in which competitive or discriminative processes intervene between perception and production (de Boer, 2001; de Boer & Zuidema, 2010).

Unbiased systematizing models easily capture majority regularization. Probability matching behavior then requires further explanation. The penchant to systematize might be weak, in relation to the statistical power of the study. In the Bayesian model proposed by Reali and Griffiths (2009), a weak systematizing prior is claimed to influence outcomes over many generations, while being poorly evidenced in the behavior of individual learners. Ashby and Maddox (1993) and Pierrehumbert et al. (2014) provide for a free parameter that controls the strength of systematization; the value of this parameter is an empirical question, and could prove to be quite moderate, so that considerable statistical power is needed to detect the nonlinearity.

Irregularization can be described in a systematizing model – or a preservatory model – by assuming noise in the learning process, such as intermittent attention to the input. The nonlinear learning rule of a systematizing model can also be combined with a substantive bias (as in Pierrehumbert et al., 2014), and depending on its strength, this bias can result in irregularizing or minority regularizing outcomes. The possibility that this bias may vary across individuals leads to an additional interpretation of reports of probability matching behavior. Apparent probability matching could arise from pooling data of participants with heterogeneous biases. This point was made for artificial language learning by Smith and Wonnacott (2010), but has been made more broadly by Gallistel, Fairhurst, and Balsam (2004) and as far back as Estes (1957). To illustrate the problem, if the input is evenly split between variant A and variant B, and $P(A) = 1.00$ in the outputs of half the participants while $P(A) = 0.00$ for the other half, the pooled data would appear to show probability matching, even though each individual produced a completely consistent system. These observations point to the importance of evaluating models on the basis of their predictions about inter-participant variability, on the assumption that inter-participant variability arises when individuals adopt different values for free parameters in the model.

2.3. Inter-participant variability

A central goal of the present study is to shed light on inter-participant variability. We investigate the extent to which individual learners differ from one another, both within and across exposure conditions. We run far more participants than other studies, in order to overcome the

difficulties of interpretation associated with low statistical power. In light of the discussion above, we will ask whether each participant deviates significantly from probability matching behavior, whether there is evidence for a substantive bias, and how distributions of participant behavior patterns compare across the conditions of the experiment.

Related work on first and second language acquisition already leads us to expect that important inter-participant variability will be found. Such variability can arise from persistent individual traits, or from situation-dependent individual states, such as differences in attention or anxiety level (Chen, Gully, Whiteman, & Kilcullen, 2000). Bates and MacWhinney (1987) and Skehan (1998) attach a prominent role to individual differences (understood as persistent cognitive traits) and relate them to different learning outcomes. Siegelman, Bogaerts, Elazar, Arciuli, and Frost (2018) present a detailed evaluation of the stability of individual differences in learning linguistic patterns. Two sociophonetic field studies (Scobbie, 2006; Stuart-Smith, Pryce, Timmins, & Gunter, 2013) find that young adults in a dialect contact situation resolve the inconsistency differently, suggesting the importance of individual socio-cognitive factors. Schmidt (2012) argues that variation amongst individuals in attention substantially impacts the outcomes of foreign language learning. Perfors (2012) found that manipulating the social assumptions in a word-formation experiment leads to overuse of the regular (e.g the dominant) form by a subgroup of participants. In an artificial language learning experiment, Rácz, Hay, and Pierrehumbert (2017) found that participants vary greatly in their success in identifying relevant linguistic and socio-linguistic cues for a word-formation pattern. For a task involving descriptions of people in different categories, Heit (1994) reports that some individuals used prior knowledge much more than others.

Inter-participant variability may also be very important for understanding language change. Blythe and Croft (2012), Fagyal et al. (2010) and Nettle (1999) outline how variability in individuals' communicative choices may influence outcomes in language change. Pierrehumbert (2012) surveys studies indicating that heterogeneity in the speech community is key to explaining empirically observed rates of change. As noted in this review, mathematical models of homogeneous speech communities converge to stable linguistic norms, which is an unrealistic outcome. In reality, languages always change over time.

Here we concentrate on inter-participant variability in learning behavior with a view to documenting its extent and its statistical properties. A question of particular interest is whether a single free parameter is sufficient to characterize the variability, or whether more than one parameter is needed. As we will show, the answer to this question can provide diagnostic information about the learning mechanism. To have the statistical power to obtain good information about the distributions of participant behavior, we run a much larger number of participants in each condition than other studies have. However, we do not have the multiple test results for each individual that would allow us to discuss individual differences in the standard meaning of the term. Our study of inter-participant variability builds directly on a smaller study by Schumacher, Pierrehumbert, and LaShell (2014). They investigated the interaction of inconsistency and number-marking system, piloting the experimental paradigm that we adopt here. They observed a strong interaction between the familiarity of the system (Plural vs. Singulative) system and the presentation frequency. Underpinning this interaction was great variability in the response patterns when the input system was inconsistent (at $P = 0.75$) and the marking system was Singulative (e.g. less familiar to English speakers.) The output for some learners was approximately probability matching, others regularized the majority pattern, and others irregularized or regularized the minority pattern. This was a significant finding. It is inconsistent with important proposals in the research literature about universal properties of language learning. Proposals holding that learning is basically preservatory, or that there is a universal bias towards the more familiar pattern, or that people regularize the statistically dominant pattern, all fail to capture the behavior of some of the participants in this study. An investigation of greater depth is needed to replicate the key findings, explore the conditions on their occurrence, and make more exact comparisons to the predictions of previous learning models.

In the present study, we expand on previous work to (i) test whether learning of inconsistency is preservatory or systematizing for a new type of input (ii) characterize biases and (iii) ascertain the nature and extent of inter-participant variability. Participants were taught an artificial language with a manipulation known as morphological reversal using a paradigm novel to the study of inconsistent systems, *unscaled adaptive tracking* (Leek, 2001). We analyze the outcome in relation to several existing learning models, and discuss possible reasons, including design limitations, for the contrast between previous findings and those of our study. Finally, we discuss the implications of our findings for the mechanisms of language change.

3. Experiment

3.1. Participants

Participants were recruited via Amazon Mechanical Turk (AMT). AMT and other on-line platforms have become widely used in psycholinguistics because they make it possible to recruit larger numbers of participants than can be brought into the lab. This was necessary for our design, because we needed a large number of subjects to characterize inter-participant variability. The quality of data obtained in this way has been validated in studies by Snow, O'Connor, Jurafsky, and Ng (2008), and Warriner, Kuperman, and Brysbaert (2013). Even though on-line data collection excludes people without Internet access, it still enables researchers to recruit a participant pool that is more diverse than the typical pool of university undergraduates (Gosling, Sandy, John, & Potter, 2010). The mean age of the participant population was 36 (σ = 10.65), and it was 57% female. An on-line platform for the experiment therefore advanced our goal of exploring inter-participant variability since we were able to collect data from a more diverse sample than is typical.

Six hundred and eighty-one (681) participants were recruited and run on AMT. Participants were paid three (3) dollars for their participation. Only participants who followed the instructions and completed the experiment were included. Ninety-six percent (96%) of participants completed the experiment, and the average time to completion was 12 minutes. Five participants who were not native speakers of English were excluded, as were three participants who reported language disabilities. Fourteen participants were then randomly eliminated until there was an equal number in each condition (79). The analysis was conducted on the remaining six hundred and thirty-two (632) participants.

3.2. Manipulation

Our study contrasts two different systems for marking number on nouns. Number marking is easily imageable, and English-speaking learners are familiar with number marking as a morphological category. In a *plural-marking* system, the bare form of the noun refers to a single occurrence of a referent, and a suffix is added if there are more occurrences. The plural system is typologically most common, and is the system used in English. In a *singulative-marking* system, this pattern is reversed. The bare form denotes multiple occurrences, and a suffix is added to denote a single occurrence. Singulative systems are rare (Anderson, 1985; Haspelmath & Karjus, 2017), but do occur in a few languages such as Welsh, Turkana, Dagaare, and Maltese (Grimm, 2012b). English does not have singulative marking, even as a minority pattern; exceptions to the dominant English plural *(e)s* pattern use a different affix for the plural (e.g. *ox, oxen*), or else a null marking (e.g. *sheep, sheep*). The contrast between English and Welsh is illustrated in Fig. 1.

Languages that have singulative-marking do not use it for every noun. Some noun stems take singulative-marking whereas others take

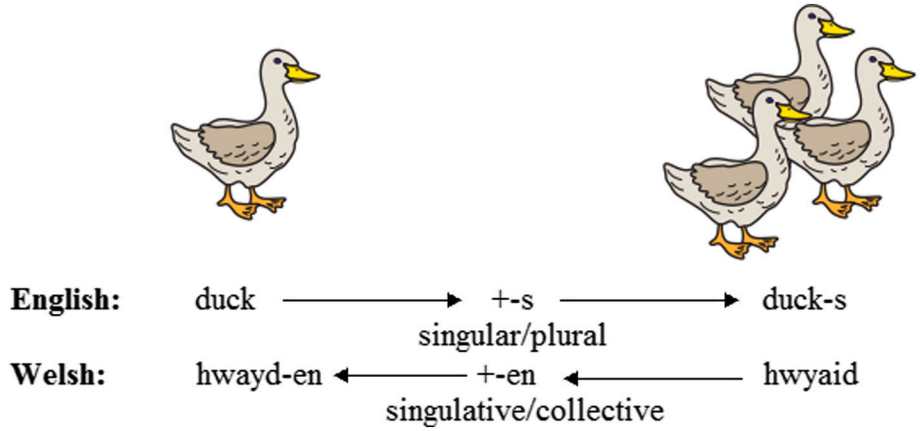

Fig. 1. Comparison of plural-marking in English and singulative-marking in Welsh.

plural-marking. Across languages, there are statistical trends in which nouns are assigned to the singulative class, since entities that typically occur in groups are more likely to be in the singulative class (Grimm, 2012b; Haspelmath & Karjus, 2017; Kurumada & Grimm, 2019). However, there is also considerable unpredictability since the assignment depends on the construal of the individuation or collectivity of the noun. In English, we can discuss either the "chairs" in a conference room (individuated) or the "seating" (as a collective). Welsh has singulative/collective *hwyad-en* "duck"/ *hwyaid* "ducks", but singular/plural *gwydd* "goose"/ *gwydd-au* "geese". Singulative-marking thus provides a very realistic example of a case in which a language learner is presented with an inconsistent system, which is eventually learned.

In our study, the same affix is used to mark the singulative on some noun stems and the plural on other noun stems. Baerman (2007) terms this situation – one in which the same affix is used to mark opposing systems – *morphological reversal*. Morphological reversal with singulative– and plural– marking does occur in certain languages, such as Masalit, where the affix *–di* encodes the plural in some cases and the singulative in other cases (Dimmendaal, 2000). A similar situation also occurs in Dagaare (Grimm, 2012a). This manipulation ensures that the imageability of the referents (single objects versus groups of objects), and the form of the affix itself, are controlled. Because singulative-marking is both typologically rare and unattested in English, it is less familiar to adult English speaking participants than plural-marking. In that way, it is similar to the typologically anomalous word sequences that Culbertson et al. (2012) manipulated, or the phonetically ungrounded vowel co-occurrence rules explored by Baer-Henney et al. (2014). Our manipulation is different, however, from studies that vary the presence or absence of an element, such as Hudson Kam and Newport (2005, 2009) or the association of some object with a set of alternative word forms (Reali & Griffiths, 2009; Vouloumanos, 2008), because we vary the level of systemic inconsistency. That is, the different experimental conditions vary the proportion of singulative-marking stems in the lexicon for the participant. But the marking system for a stem (e.g. its inflectional class) is not randomly varied over the course of the training.

There were 8 conditions organized in a 2 × 4 design. The first factor is the dominant marking system, Plural or Singulative. The second factor, consistency, is coded by the frequency of the dominant pattern in the training set, expressed as a probability. The four consistencies were: 1.0, 0.875, 0.75 and 0.625.

3.3. Procedure

3.3.1. Structure of the task

The experiment included a training phase and a test phase. Training trials were two-alternative forced-choice with immediate feedback. The test phase included generalization trials as well as all training trials repeated to assess recall of training items. Test trials were two-alternative forced choice with no feedback.

3.3.2. Unscaled adaptive tracking

Training was conducted using a modification of the *adaptive tracking* paradigm. Adaptive tracking, also known as Bekesy tracking, is a technique used in audiology (Leek, 2001). In adaptive tracking, participants progress through an ordered series of trials, advancing to the next trial in the series by providing a correct answer to the current trial. The participant regresses to the previous trial if an incorrect response is given. Regression to the previous trial takes place regardless of whether or not a correct answer had previously been provided. That is, if a participant regresses from trial *t* to trial *t-1* and then provides an incorrect answer at *t-1,* the participant would then regress to trial *t-2*. The participant would then have to provide correct answers again to *t-2*, *t-1* in order to return to *t*.

In traditional adaptive tracking, trials increase in difficulty as the task progresses, however the stimuli in our task were not graduated in difficulty (the task is thus *unscaled*). For each trial, the participant provided a response to a particular stimulus. They received feedback about the correctness or incorrectness of their response in the form of a visual display that showed whether they were progressing or regressing.

Adaptive tracking is commonly employed in computer games, and the task was presented to participants as a computer game. Because of its similarity to games that people play for fun, we expected it would focus participants on the task. In this paradigm, the length of training is not defined by a fixed set of exposures. Participants can complete training quickly once they have learned the system. For example, if a participant correctly infers that an affix is always singulative-marking, that participant can proceed through training without making any more mistakes, thereby completing it quickly.

We view this feature of the paradigm as an advantage for addressing our research questions. In many contemporary artificial language learning experiments, participants are forced to go through dozens or even hundreds of trials over the course of several days. Such training may continue well after the participant has learned the input. By keeping fatigue to a minimum, unscaled adaptive tracking may also allow participants to be more engaged for the test phase of the experiment. Lastly, the paradigm facilitates detection of inter-participant variability. Designs with fixed-length training may not present enough trials for some participants to orient to the task or overcome learning difficulties, leading to poor performance in the test phrase. Although such participants would be excluded in a standard analysis, excluding them may effectively eliminate legitimate records of inter-participant variability. By training until participants have responded correctly to all the stimuli, the paradigm minimizes the exclusion rate and improves

the researcher's ability to characterize the full range of variation. Detailed records of performance during training also provide information about the learning process that is not available in paradigms with fixed-length training.

This kind of task may encourage more explicit hypotheses from participants than the implicit learning tasks in artificial language learning studies like Hudson Kam and Newport (2005, 2009) for two reasons. In the first place, as we mentioned, participants are incentivized to posit rules that will allow them to complete the task quickly. Additionally, the task provides learners with feedback. Feedback is a form of negative evidence, and the use of negative evidence by language learners is controversial (Marcus, 1993). This design choice may by itself cause the results to diverge from previous studies that have not used feedback. However, there is significant value in extending existing theories to novel tasks with different types of exposure. Feedback is one of those circumstances, since adult language learning can involve explicit learning (Chouinard & Clark, 2003, Ellis, 2015; Hulstijn, 2015) and there is even evidence that children make use of adult corrections to infer the bounds of grammaticality (Saxton, 2000; Saxton, Backley, & Gallaway, 2005).

To make the task seem more like a game to participants, a brief storyline was provided. Participants were told that they would have to cross a body of water to reach a castle. The castle belongs to the fairy "Bendith", and to cross the body of water they would have to guess what the words for certain objects were in "fairy language" (Fig. 2). If they answered correctly, the fairy would reward them by providing a plank for the bridge that allows them to cross the water. If they answered incorrectly, the fairy would get angry and break the last plank which had been placed down, and the player would regress to the last stable bridge plank (Fig. 3).

At each section of the bridge, participants were shown the trial for that section and shown both affixed and unaffixed forms on buttons above the image. Participants were instructed to click on the button with the word that they thought was correct. During the test phase, the player is shown standing in front of the portcullis of the castle, and must provide an answer for each test trial without receiving any feedback before the portcullis is raised. Completing the test phase was required in order for the participant to be paid.

3.4. Materials

3.4.1. The input language

The input language consisted of a list of lemmas that pair a stem with a referent. The stems were five characters long, and did not correspond to any English word. They were built using bigram statistics drawn from the *Cronfa Electroneg o Gymraeg* ("Electronic Corpus of Welsh"; see Ellis, O'Dochartaigh, Hicks, Morgan, & Laporte, 2001) to make the stems look sufficiently distinct from English to show the participants that they were not seeing English words. This was done to minimize any possible influence of specific regular or irregular plural forms in English. Each lemma was in turn associated with two word forms, which are the two inflectional variants of the lemma. In one, the stem was bare, and in the other, it was followed by the suffix *-yl*, which also has a Welsh appearance and does not correspond to any suffix of English. There were 32 distinct stems, and thus 64 distinct word forms, which are listed in Appendix A. The 32 referents were drawn as images by an artist; they are also shown in Appendix A. 16 were used for training and 16 were reserved for the test phase, as indicated. For each type, there was both a single-token image and a multiple-token image (showing five tokens of the referent). Thus, there were also 64 distinct images of word form referents. Some of the referents, such as the sheep and the side-chair, more typically occur in groups than others, such as the bear and the wrench. However, all of them could plausibly be found either singly or in a group.

The input language had a morphological class system. If the dominant marking system was Plural, then most or all of the lemmas used the bare form for a single token of the referent, and the suffixed form for multiple tokens. In the inconsistent conditions, the language also had a minority of lemmas for which this situation was reversed; the bare form referred to multiple tokens, while the suffixed form referred to the single token. If the dominant marking system was Singulative, then the most or all of the lemmas used the bare form for multiple tokens and the suffixed form for a single token; the minority-pattern lemmas had the reversed marking system. For all conditions, the participants encountered both forms of every lemma during the course of the experiment.

Morphological reversal in the input language is a unique manipulation in artificial language learning experiments. It combines lexically conditioned variation like that in Wonnacott and Newport (2005), Wonnacott, Newport, and Tanenhaus (2008), with the more typical unconditioned variation in studies like Hudson Kam and Newport (2005), Reali and Griffiths (2009) and Culbertson et al. (2012). Any

Fig. 2. A training phase trial. The player is the mushroom (left), standing on the plank set down from the previous successful trial. In this trial, the participant sees the reference objects (baskets) and has to choose which of the two words "wiben" or "wibenyl" is correct. If the participant chooses the correct word, then a plank is added to the pillars and the player advances.

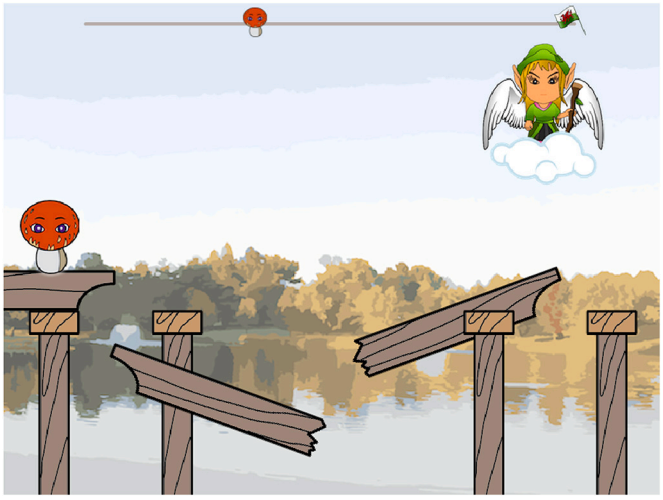

Fig. 3. The consequence of an incorrect response. The animation shows the mushroom hopping back to the previous stable plank, where it is situated in this picture, and then Bendith breaks the unoccupied plank.

given lemma always used the same marking system, meaning that the marking system associated with the affix was lexically conditioned. But the assignment of lemmas to morphological classes was arbitrary. No property of either the stem or the referent image gave information about the marking system for that particular lemma. Therefore, the marking system associated with the affix, or its absence, was unpredictable in the sense used by Hudson Kam and Newport (2005) and Reali and Griffiths (2009) until (or unless) the assignment of the stem to marking system was learned. To guess the correct form on a trial that was seen for the first time, the learner had two potential sources of information. The most important was the overall consistency for the dominant system. In addition, if the participant were able to remember their answer for one of the inflected forms of a lemma, they might have used that information to select the complementary inflected form for the complementary referent when it was presented later. The results in the training phase are tabulated by trial (rather than by lemma) since feedback was provided at each trial. For the test phase, almost a fifth of all responses on the test phase (19%) had both forms for a lemma either marked or unmarked. Out of all the lemmas that a participant did not assign to the dominant system, 65% had either both forms marked or both forms unmarked; only 35% displayed complementation between a marked form and an unmarked form. This situation suggests that participants were not effectively using their previous classification for subsequent classifications. This is not surprising since the participants saw each novel item in the test phrase only once. Going forward, all results will therefore be tabulated by word form and not by lemma.

The artificial language in this experiment is simple compared to other studies like Hudson Kam and Newport (2005, 2009), which used full sentences with a larger vocabulary. While sentence-level input may provide tangible benefits in some circumstances, we chose to use just nominal-level input because it would focus participants on just the relevant dimension (see Perfors (2012, 2016) for a similar argument).

3.5. Randomization and presentation

3.5.1. Training phase

Recall that the study has a 2 × 4, between-subjects design. The two systems (Singulative, Plural) were taught to different groups of participants at each of four different consistencies (1.0, 0.875, 0.75, 0.625), for a total of 8 experimental conditions.

For each participant, a fresh instance of the language system was created by randomizing with regard to multiple factors. First, a fresh random assignment of stems to referents was made, within the training set and within the test set. The training lemmas were allocated at random to morphological classes, according to the frequencies for each consistency condition. Thus, the 1.0 conditions had no minority-pattern lemmas; in the 0.875 condition, there were 2 such lemmas, in the 0.75 condition, there were 4, and in the 0.625 condition, there were 6. The word forms for lemmas in the novel test set did not need to be assigned to a morphological class, because no feedback was provided about the correct answer.

As already shown in Fig. 2, on each trial a single-token or multiple-token image was displayed with a choice between the two forms of the lemma. If the lemma belonged to the dominant morphological class, then the correct answer was the word form reflecting the dominant marking system; but if (unbeknownst to the participant), the lemma belonged to the minority morphological class, then the correct answer was the opposite. The sequence of 32 trials was block-randomized into 4 blocks of 8, counterbalancing along two dimensions: (a) the number of times the word form bearing the suffix was the correct answer and (b) the proportion of dominant-class trials versus minority-class trials. Thus, in the 0.875 conditions, the 2 minority-pattern lemmas define 4 minority-pattern word forms, and one of these was assigned to each of the 4 blocks. The 0.75 conditions had 2 minority-pattern word forms per block, while the 0.625 conditions had 3 per block. Counterbalancing on the number of tokens in the image would also have been desirable. For the 1.0 and 0.75 conditions, each block did contain 4 single-token trials and 4 multiple-token trials. For the 0.875 and 0.625 conditions, this goal could not be perfectly achieved while meeting the other requirements, and each block contained either 5 or 3 single-token images. Each block was randomized, which means that the minority-pattern trials occurred at unpredictable locations. Example training blocks for all conditions are shown in Appendix B. Note that both forms of a lemma may appear in the same block, but this happens only sporadically.

The training phase of 32 (unique) trials was short compared to many artificial language learning experiments (Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Hudson Kam & Newport, 2005, 2009; Culbertson et al., 2012; Perfors, 2016, inter alia). However, it had a larger number of distinct word forms and referents than many other studies (Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Vouloumanos, 2008). Although the training phase was relatively short, unscaled adaptive tracking aids in learning the correct input probabilities by requiring correct categorization of every trial. The input was also comparatively simple; unlike the Hudson Kam and Newport (2005, 2009) studies, for example, there was no articulated grammar to be learned. The only characteristic of the input that had to be learned on each trial was the association between the presence/absence of the affix and the number of referents. In this way, the simplicity of the input in conjunction with the advantages of the training paradigm mitigated the

impact of a short training phase.

3.5.2. Test phase

The test phase included both the 16 seen stems and the 16 novel test stems. Thus, it included both inflectional forms of all 32 lemmas, for a total of 64 trials. These trials were randomized without blocking for each participant.

3.6. Hypotheses

To elucidate the predictions of preservatory versus systematizing mechanisms, we consider generalized versions of influential preservatory and systematizing models. We take models whose output probabilities converge to input probabilities, in the absence of bias, to be preservatory. By contrast, models that converge to regularization are basically systematizing. The first model we consider is the learning model of Estes (1957), building on Bush and Mosteller (1951), sometimes known as *linear reward penalty*. The Estes-Bush-Mosteller model is one of the most influential learning models in all of mathematical psychology. It has been applied to the theory of language learning in Yang (2005) and to the question of regularization in artificial language learning by Rische (2014) and Ma and Komarova (2017). The model's historic influence and broad applicability (Gallistel, 1990) make it a highly relevant exemplar of preservatory learning via either implicit feedback (Rische, 2014) or explicit feedback (Paul & Ashby, 2013).

The model details are described briefly below. Taking m_i to be the mental estimate of the frequency at time i and s_i to be the observed frequency in the input at time i, learning is described by a single free parameter θ, representing a learning rate. The updated mental estimate is a weighted average of the previous mental estimate and the input frequency, as shown in the following equation:

$$m_{i+1} = \theta(m_i) + (1-\theta)s_i \tag{1}$$

We will use a high m_i to represent a high expectation for the Plural system at the start of the experiment. If the initial expectation is coupled with a high value of θ, the individual persists in their initial belief, placing very little weight on the input. However, with a low value of θ, the participant readily adjusts their mental state to the new input. In that case, the model will converge quickly to the recently observed frequency.[2]

Another highly influential class of models that have been applied to artificial language learning are Bayesian models – particularly the beta-binomial. Because these models are fundamentally preservatory, they generate regularization only via sampling effects or substantive biases. A particular form of the beta-binomial Bayesian model was advanced in Reali and Griffiths (2009) as a neutral model of language change. The authors propose that it causes regularization over long periods of time in all cases where no substantive or functional factor interferes. It has been used to explain regularization in the non-linguistic domain as well (Ferdinand, Thompson, Kirby, & Smith, 2013). It was extended by Culbertson and Smolensky (2012) to cover selection of particular variants.

Because the Culbertson-Smolensky experiment manipulated two linguistic dimensions, their beta-binomial Bayesian model has two distinct dimensions and mixing weights. We simplify their model to a single dimension here, in order to apply it to our experiment. The prior of the beta-binomial has two hyperparameters α, β, the sum of which is greater than 0, and which together determine the shape of the distribution. An important difference from the Estes-Bush-Mosteller model is that the prior is not a single probability, but rather a distribution over probabilities. If $\alpha = \beta$, the distribution is symmetric and there is no preference for either of the two competitors. The relation of α to β, as reflected in the ratio $\alpha/(\alpha+\beta)$, represents the overall extent of the bias for one competitor over the other. Here, we will take the case of $\alpha/(\alpha+\beta) > 0.5$ to represent bias towards the Plural. The posterior after training is also a beta distribution, in which the values of α and β have been updated by the counts of the outcomes observed in the training. Taking A to be the number of examples of Plural in the training and B to be the number of examples of Singulative observed in training, the expected value of the posterior after training is:

$$\frac{\alpha + A}{\alpha + A + \beta + B} \tag{2}$$

This equation leads to the observation that the initial values of α and β can be intuitively understood as counts of examples. For a conservative learner who is little influenced by the training, $\alpha + \beta$ would be very large (in relation to the number of stimuli in the experiment). For a very adaptable learner, $\alpha + \beta$ would be much smaller.

The Estes-Bush-Mostellar model and the beta-binomial Bayesian model share some important similarities. First, the prior assumption is strongly reflected after small amounts of input, and more weakly after large amounts. Therefore, a learner exposed only to a sparse sample from a language may produce outputs that reflect the prior. Second, both provide a way to parameterize the strength of the prior. In the Estes-Bush-Mosteller model, θ represents the extent to which the individual perseverates in a previous belief. For the beta-binomial Bayesian model, the strength of the prior is represented by the effective total count of examples $\alpha + \beta$ that the learner has in mind at the start of the experiment. This is much smaller than the number of examples previously encountered in a learner's experience, because the experiment is short in comparison to the age of the participants. If the effective count is much larger than number of stimuli presented, then it will dominate the outcomes, whereas if it is much smaller, then the stimulus patterns will dominate the outcomes.

The similarity of these models can be brought out by calculating the range of potential outcomes as the strength of the preference for the Plural system is varied, for the exact conditions in our study. For these calculations, we presuppose 32 unique training items at each consistency. Since the participants were native speakers of English, and the English Plural system is also the typologically dominant system, we assume prior experience favors the Plural, and that the open question is the extent to which this prior experience influences learning of the current task. For the Estes-Bush-Mosteller model, we assume that initially $m_0 = 1$, and θ is varied between *0.5* (rapid learning) and *1.0* (no learning). For the beta-binomial Bayesian model, we fixed $\beta = 1$ and varied α from 1 to 2000.[3] $\alpha = \beta = 1$ describes a uniform prior (previous experience with the Plural has no influence on the task), and as α is increased, the influence of the previous experience also increases. Fig. 4 displays the range of possible outcomes predicted for each condition in our study, based on these assumptions.

The blue lines represent the no-bias scenario, where both models are

[2] In our calculations, we take the "observed frequency at time i" to be the frequency observed in the correct answers for the set of all stimuli observed up to time i. If we were to assume instead that the observation interval is only the most recent trial, then s_i reduces to a Boolean variable (e.g $s_i = 1$ or $s_i = 0$.) For θ=0.0, the mental state would just match the most recent example, without tracking the longer-term statistics. For any θ<0.5, the most recent trial would dominate the mental state. This scenario is inconsistent with results on training presented above, which indicate that participants improved over the course of the training.

[3] Reali and Griffiths (2009) work with a model in a very different parameter range for the beta distribution, $\alpha = \beta < 1$. These parameters provide a symmetric U-shaped prior that favors systematizing without favoring either competitor. We do not present calculations with this model because it predicts negligible individual variation after 32 distinct trials. Reali and Griffiths indeed claim that the effect of the prior is very slight, only becoming evident as learning is iterated over many generations, contrary to the outcomes we will report below.

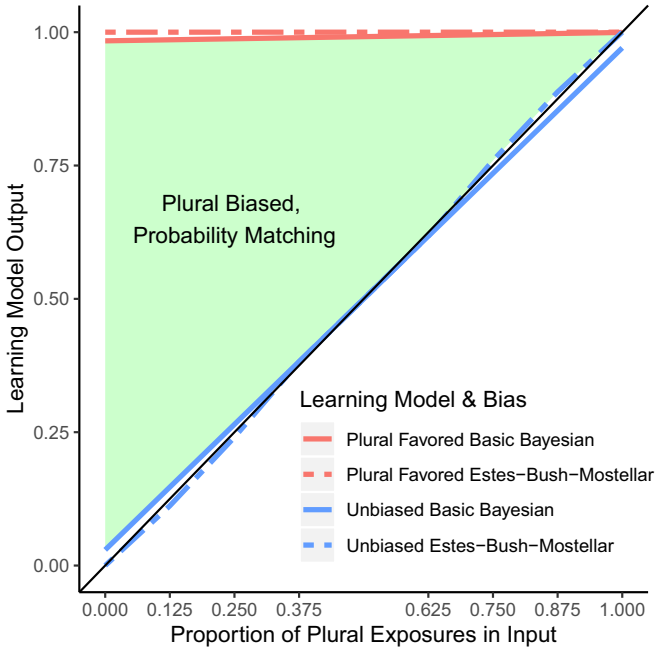

Fig. 4. Predictions generated for each condition, represented as the absolute proportion of plural forms in the input. The colored lines indicate the model and circumstance as shown. The black line $y = x$ represents exact probability matching. The plot was generated with initial parameters $\theta = 1$, $m = 1$ for the biased Estes-Bush-Mostellar model and $\theta = 0.5$, $m = 1$ for the unbiased model. The biased beta-binomial Bayesian model used hyperparameters $\alpha = 2000$ (but effectively, $\alpha \to +\infty$), $\beta = 1$, and the unbiased version used neutral hyperparameters $\alpha = 1$, $\beta = 1$.

roughly equivalent for the given starting parameters and generate outcomes that differ very little from probability matching. The lines are slightly different because slight influences of $m_0 = 1$ and $\beta = 1$ still remain after 32 training items. Values in the green shaded area represent possible directional shifts for biased preservatory learning that can be captured under either model. Both models can produce any value above probability matching through extreme initial states. The large lower-triangular region in the figure below the blue lines shows what should *not* occur. Since the space of possible outcomes is upper-triangular, both models predict that irregularization and minority regularization will not be found in the four Plural dominant input conditions. These outcomes could only arise through other factors, such as noise or sparse sampling.

We turn now to the predictions of a purely systematizing mechanism, without a substantive bias. Such a mechanism strives only to make the output more regular than the input, which is achieved by increasing productions of the dominant system at the expense of productions of the alternative. Such a preference can be modelled by essentially any sigmoid function that passes through (0,0), (0.5, 0.5) and (1,1) (for specific

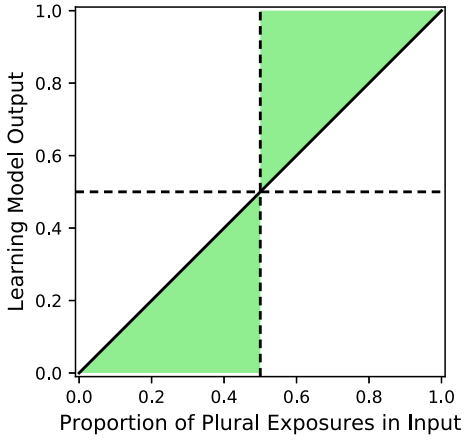

Fig. 5. A plot of the expectations for an unbiased systematizing mechanism. The shaded region depicts the where outputs are expected, given the input proportion on the x-axis. For proportions above the dashed reference lines at 0.5, one variant is a majority and thus the optimal variant for regularization.

proposals, see Ashby & Maddox, 1993; Kirby et al., 2007; Mandelshtam & Komarova, 2014; Pierrehumbert et al., 2014). The strength of the systematizing mechanism can be summarized by the slope at (0.5, 0.5); as the slope approaches infinity, the function approaches the statistically optimal threshold decision rule, by which the majority variant is produced all the time. Thus, in Fig. 5, the shaded region shows where outputs could fall for the different conditions of the study.

Note that if the mechanism is unbiased and systematizing, irregularization and minority regularization are not predicted outcomes. These outcomes could only arise as a result of noise or sparse sampling. Furthermore, the space of outcomes for the Plural dominant systems is essentially identical in Figs. 4 and 5. For the Singulative dominant systems, in contrast, the predictions are completely different; in Fig. 4 the shaded region for $0 < x < 0.5$ is above the line $x = y$, whereas in Fig. 5, it is below the line.

We have also posed the question of whether substantive biases and propensities to systematize are universal, or whether the strengths of such preferences are specific to the participant. If they are universal, outcomes for all participants are predicted to cohere together, somewhere within the shaded area of Figs. 4 or 5. If they are specific to the participant, however, Figs. 4 and 5 receive fresh interpretations, on the assumption that the participants from every condition provide a representative sample from the range of outcomes generated by the equations of the model. On this assumption, responses by different participants would be spread out, and the shaded region of each figure shows where they can spread out. For preservatory learning, the upper-triangular shape in Fig. 4 means that the variability amongst participants should be greatest in the Singulative 1.00 condition, and decrease monotonically as the proportion of Plural items increases. The reason for this is ceiling effects; any variation in the strength of a Plural bias has space to express itself in the Singulative 1.00, while this ability decreases as the proportion of the Plural increases. For unbiased systematizing learning, the predictions are very different. As shown in Fig. 5, there is most room for variability in the most inconsistent conditions (0.375 and 0.625), and the variability should decrease if the input is more consistent.

Finally, it is possible that the learning mechanism is both biased and systematizing. Since the predictions in Figs. 4 and 5 are consistent when the input is Plural dominant (the systematizing preference is aligned with the substantive bias), various mathematical formulations (notably including that of Pierrehumbert et al., 2014) would predict regularization. When the input is Singulative dominant, however, the structural and substantive biases are not aligned; in fact, the predictions of the preservatory and unbiased systematizing models do not overlap at all. The union of the mutually exclusive predictions for Plural rates under 0.5 in the two Figures encompasses all outcome probabilities from 0 to 1.0. This indicates that when a substantive bias and a systematizing preference are in conflict, their interaction is an open question. Obtaining an empirical characterization of this interaction was a central goal in the design of the present study.

3.7. Results

In this section, we present both descriptive statistics that display the main patterns in the data, and the results of statistical tests to determine the significance of the factors manipulated. In the training phase, there are two measures of interest: the number of attempts required to complete the training phase ("steps") reveals the difficulty of the task, and the proportion correct for each training block reveals the extent to which participants succeeded in improving their guesses about the marking system as training progresses. The measure of interest in the test phase is the proportion of responses consistent with the dominant marking system. In particular, performance on novel test items reveals how participants generalized from the training items, and then, by inference, how they processed and encoded the input.

Table 2
Average steps to completion for each condition.

Condition	Mean
Plural 1.00	39
Singulative 1.00	48
Plural 0.875	48
Singulative 0.875	59
Plural 0.75	51
Singulative 0.75	67
Plural 0.625	67
Singulative 0.625	67

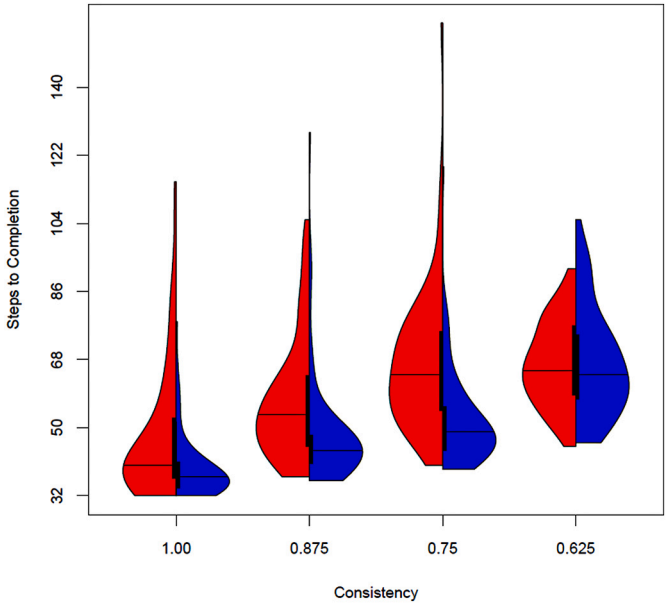

Fig. 6. Steps to completion in the training phase. Presentation consistency is on the x-axis. Singulative conditions are in red, Plural conditions in blue. The horizontal line is the median. The black bar shows the interquartile range. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.7.1. Training phase

Table 2 gives the number of steps required to finish the training phase for each condition.

The violin plots in Fig. 6 give distributional information about the number of steps required to finish in each condition.

Less consistent systems required longer to complete the training. Singulative conditions took longer to complete than Plural ones, apart from the 0.625 consistency where the completion times are nearly the same. Table 3 shows relevant comparisons using Wilcoxon rank-sum tests to compare the number of steps required in training.

The Bonferroni-corrected significance level for P in the table is 0.00625. All of the comparisons are significant, except the comparison between the Plural and Singulative 0.625.

The proportion of correct first responses produced at each training block is represented in Fig. 7 below. The horizontal axis of the plot is the training phase trials presented as blocks.

The proportion of correct responses on the training phase were analyzed using mixed logistic regression with a logit link function to assess the degree to which participants improved over the course of training. The model included random effects of participant and referent image and fixed effects of marking system, consistency, trial number, and lag-1 response. The analysis was conducted in R using the lme4

Table 3
Wilcoxon-rank sum test results for each comparison.

Group 1	Group 2	W	P
Plural	Singulative	34,578	<0.001
1.00	0.875	5946.5	<0.001
0.875	0.75	8588.5	<0.001
0.75	0.625	7924	<0.001
Plural 1.00	Singulative 1.00	2037.5	<0.001
Plural 0.875	Singulative 0.875	1646	<0.001
Plural 0.75	Singulative 0.75	1236	<0.001
Plural 0.625	Singulative 0.625	2982.5	0.6324

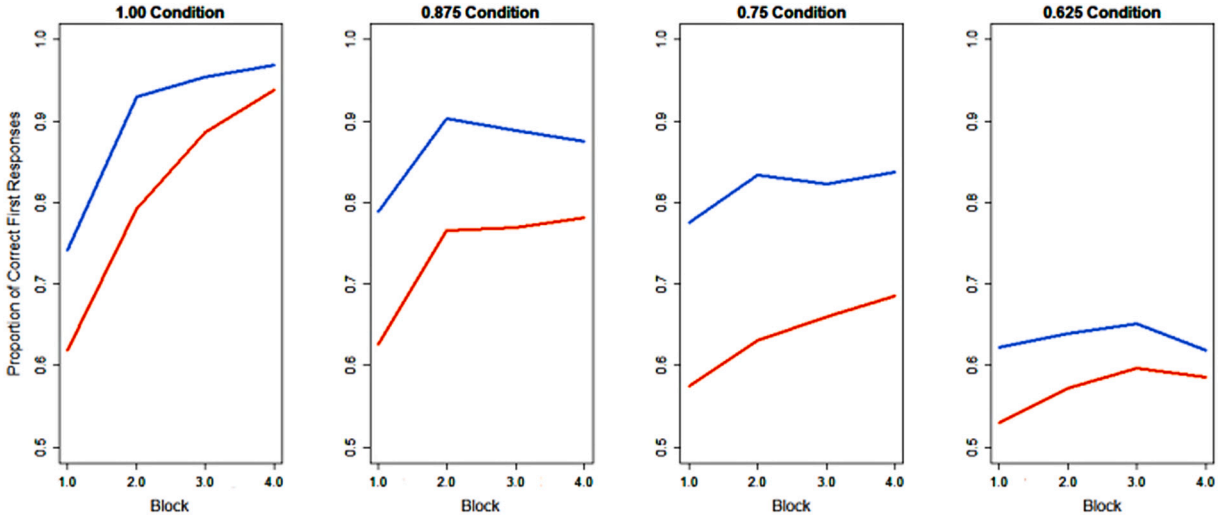

Fig. 7. Proportion of correct first responses for each training block (y-axis) in each condition. Singulative conditions are in red, Plural conditions in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Mixed Logistic Regression (Logit Link) model results on training trials.

Effect	B	Z score	P
Intercept (Reference: Plural)	0.14	1.17	0.24
Singulative	−0.53	−9.48	<0.001
Consistency - Linear	−1.35	−15.35	<0.001
Consistency - Quadratic	0.02	0.18	0.85
Consistency - Cubic	−0.23	−2.85	0.004
Trial number	0.06	21.97	<0.001
Lag-1 response	0.46	5.14	<0.001
Singulative: Consistency - Linear	0.48	4.18	<0.001
Singulative: Consistency - Quadratic	0.21	1.89	0.05
Singulative: Consistency - Cubic	0.29	2.65	0.008

Table 5
Type-III ANOVA on model of training trials.

Effect	χ^2	Df	P
Intercept	1.37	1	0.24
System	89.78	1	<0.001
Consistency	249.55	3	<0.001
Trial number	482.69	1	<0.001
Lag-1 response	26.42	1	<0.001
System:Consistency	29.76	3	<0.001

package (Bates, Maechler, Bolker, & Walker, 2015) in concert with the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2016) for significance levels. The system was coded with treatment coding, with a reference level of Plural. Consistency was coded in descending order (1.00, 0.875, 0.75, and 0.625) as an ordered factor with polynomial contrasts, which is appropriate since the intervals are equally spaced. Lag-1 response was included to account for lag-1 autocorrelations (see Baayen and Milin (2010) for discussion). The model also included a random slope of trial number within the random effect of participant to account for different learning rates. The model was run on responses to first exposures only: responses when an item is seen again after the participant has responded incorrectly are not independent from previous responses to the same item. Trial number, encoded as a numeric, corresponded to the quantity of information that had been presented to the participant, answering the question of how much that amount of information helped participants provide correct answers.[4] The final model for correct responses on the training phase found significant effects of marking system, consistency, trial, lag-1 response, and the interaction of system and consistency. Table 4 is the final model; Table 5 is a Type-III ANOVA table on the final model calculated using Anova() in R.

The effect of marking system indicates that during the training phase, participants in the Plural conditions produced more Plural responses than participants in the Singulative conditions produced Singulative responses. The main effect of consistency indicates that more consistent conditions produced more correct responses, and the effect of trial indicates that participants increased their proportion of consistent responses as the training progressed.

3.7.2. Test phase

Test phase items repeated from the training phase (*seen* items) have a correct answer, namely the same response that was needed to advance

[4] See Appendix C for the specifications of the models presented in this paper in R syntax.

Table 6
Average proportion of consistent responses, with the difference from input.

Condition	Generalization Trials		Seen Trials	
	Mean	Diff. from Input	Mean	Diff. from Input
Plural 1.00	0.97	−0.03	0.97	−0.03
Singulative 1.00	0.90	−0.10	0.90	−0.10
Plural 0.875	0.93	0.06	0.85	−0.025
Singulative 0.875	0.79	−0.09	0.75	−0.135
Plural 0.75	0.90	0.15	0.75	0
Singulative 0.75	0.67	−0.08	0.63	−0.125
Plural 0.625	0.70	0.075	0.62	−0.005
Singulative 0.625	0.54	−0.085	0.57	−0.055

during training. Items that were not presented to players during training (*novel* items) constitute generalization trials. For inconsistent conditions, a novel item has no correct classification. Generalization performance is therefore evaluated based on the number of dominant system (*consistent*) responses on a per-item basis. Table 6 summarizes the average proportion of consistent responses produced by condition for both generalization and seen trials. The mean column shows the mean production of the dominant system by condition and the difference from input column shows the difference from the input proportion (i.e, direct probability matching.)

Some conditions were close to probability matching, on the average, but deviations are found in both directions. A positive deviation

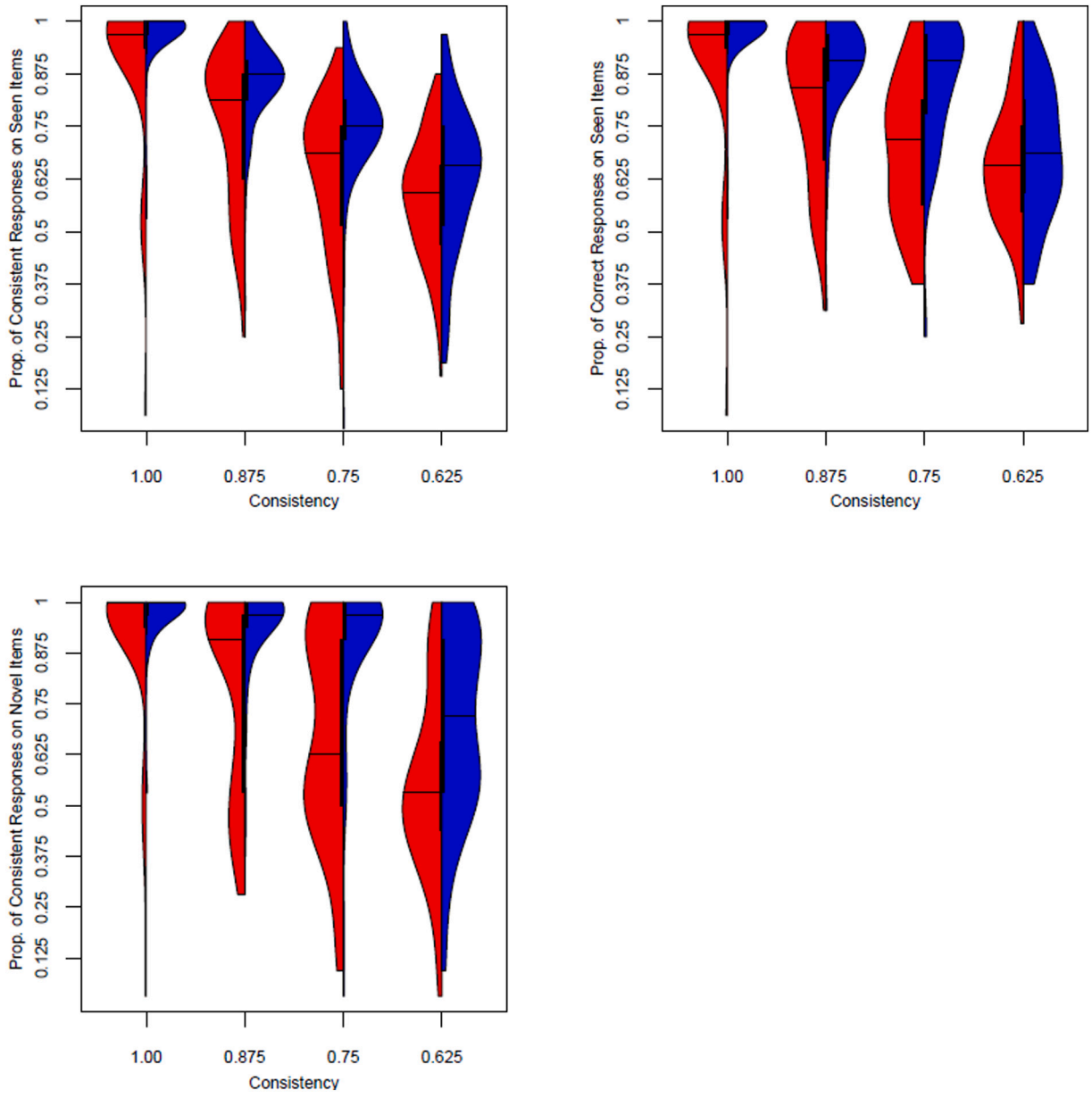

Fig. 8. Distributions of the proportion of consistent responses on seen items (upper left) and novel items (lower left). Distributions of the proportion of correct responses on seen items (upper right). Presentation consistency or correctness is on the y-axis. Singulative conditions are in red, Plural conditions in blue. The horizontal line is the median. The black bar is the interquartile range. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
Mixed Logistic Regression (with Logit Link) model results on seen trials in the test phase.

Effect	B	Z score	P
Intercept (Reference: Plural)	2.12	14.51	<0.001
Singulative	−0.72	−7.71	<0.001
Consistency - Linear	−2.78	−18.28	<0.001
Consistency - Quadratic	0.84	6.05	<0.001
Consistency - Cubic	−0.34	−2.66	0.007
Singulative: Consistency - Linear	0.70	3.54	<0.001
Singulative: Consistency - Quadratic	−0.09	−0.48	0.63
Singulative: Consistency - Cubic	0.22	1.265	0.21

represents dominant regularization, and a negative deviation is this table represents irregularization, as no condition yielded minority regularization in the averaged data. Positive deviations occur only in the inconsistent Plural conditions.

Distributions of consistent responses (the proportion of items produced consistent with the dominant system) for seen and novel items are displayed in Fig. 8; the distributions of correct responses (correct classification for a particular item) for seen (training) items are also shown for comparison.

For both Plural and Singulative conditions, participants in inconsistent conditions were far more variable in their responses than participants in the 1.00 conditions, which were near ceiling for both conditions. For the inconsistent conditions, novel items (lower left) elicited more consistent responses than seen items (upper left). This means that many participants had some success in remembering which specific forms were exceptions to the dominant pattern, since fewer consistent responses on seen items indicates more application of the minority system. This success is also reflected in the comparison between the top two panels. For example, a participant in the Plural 0.75 condition who formed a probability-matching generalization, and applied it to all the seen items, would achieve only 62.5% correct; most participants achieved higher levels of correctness in this condition.

Overall, the Plural conditions elicited more consistent responses than the Singulative conditions. However, the effect of marking system is far from constant. For the 0.625, 0.75, and 0.875 conditions, the difference is much greater for the novel items than for the seen items. Corroborating the averages in Table 6, the single most salient difference is that between the distributions of Singulative and Plural novel items in the 0.75 condition. There is more overlap between the distributions for the Singulative and Plural 0.625 on novel items, and the distributions appear quite similar for the seen items in the 0.625 as well. In summary, the effect of marking system does not appear to have a uniform interaction with consistency.

Table 7 summarizes the final mixed effects model for the seen items. It tested effects of marking system and consistency with random effects for participants and referent objects with the same method as the model reported in Table 6.

Per the ANOVA results reported in Table 8 below, there was an effect of marking system. Participants in Singulative conditions produced fewer consistent responses. The significant effect of consistency shows that participants in inconsistent conditions produced fewer consistent responses as consistency decreased. The expected interaction between system and consistency was also significant, indicating that the effect of consistency differs between Singulative and Plural groups.

Table 8
Type-III ANOVA on model of seen trials in the test phase.

Effect	χ^2	Df	P
Intercept	210.46	1	<0.001
System	59.36	1	<0.001
Consistency	334.44	3	<0.001
System:Consistency	13.58	3	0.004

Table 9
Mixed Logistic Regression (with Logit Link) results for the proportion of consistent responses on generalization trials.

Effect	B	Z score	P
Intercept (Reference: Plural)	3.32	26.59	<0.001
Singulative	−1.50	−9.39	<0.001
Consistency - Linear	−2.54	−10.61	<0.001
Consistency - Quadratic	−0.57	−2.42	0.02
Consistency - Cubic	−0.43	−1.84	0.06
Singulative: Consistency - Linear	0.01	0.05	0.96
Singulative: Consistency - Quadratic	0.99	3.10	0.002
Singulative: Consistency - Cubic	0.32	1.02	0.31

Table 10
Type-III ANOVA on model on generalization trials in the test phase.

Effect	χ^2	Df	P
Intercept	707.52	1	<0.001
System	88.13	1	<0.001
Consistency	135.49	3	<0.001
System:Consistency	11.56	3	0.009

We now turn to the regression for the generalization trials. We tested effects of marking system and consistency on the consistent responses on generalization trials, with a random effect of referent object and participant.[5] The method was the same as for the preceding models. The results are presented in Tables 9 and 10 below.

There were significant effects of marking system and consistency; Plural conditions produced more consistent responses than Singulative, and more consistent conditions produced more consistent responses than less consistent conditions. The interaction between system and consistency is again significant.

To understand the results in relation to the prior literature on probability matching and regularization in artificial language learning, it will be useful to classify participants according to the difference between the input consistency and the proportion of consistent responses produced on the novel items in the test phase. Fig. 9 displays the distributions of these differences. To aid in interpreting these distributions, two comparisons are provided. Ninety-five percent (95%) confidence intervals for outcomes of probability matching are indicated with dashed lines. Ninety-five percent (95%) confidence intervals for random guessing (0.5) between two equally likely alternatives are indicated with a green backdrop.

There is significant variation amongst participants. The number of majority regularizers depends strongly on the dominant marking system and on the presentation consistency. A large effect of marking system is found in the 0.875 and 0.75 conditions. The bulk of participants in the more familiar Plural condition fall above the probability matching dashed line in these conditions, whereas the Singulative participants are spread out on both sides of it. Several of the distributions appear bimodal. Note that none of these have a mode at zero, which represents probability matching. Rather, the lower modes fall within the green region for random guessing. These observations cast doubt on the possibility that the results are due to deviations from a preservatory

[5] One reviewer raised the concern that steps to completion in the training phase should also be considered as a factor that may have contributed to the results. Because the steps to completion differ across the training conditions (Table 2), participants might have become more fatigued in some conditions than in others, affecting their performance during the test phase. We carried out a separate analysis to evaluate this possibility. We found that steps to completion in the training phrase is a very weak predictor of regularization behavior, and cannot explain the large differences found across the conditions.

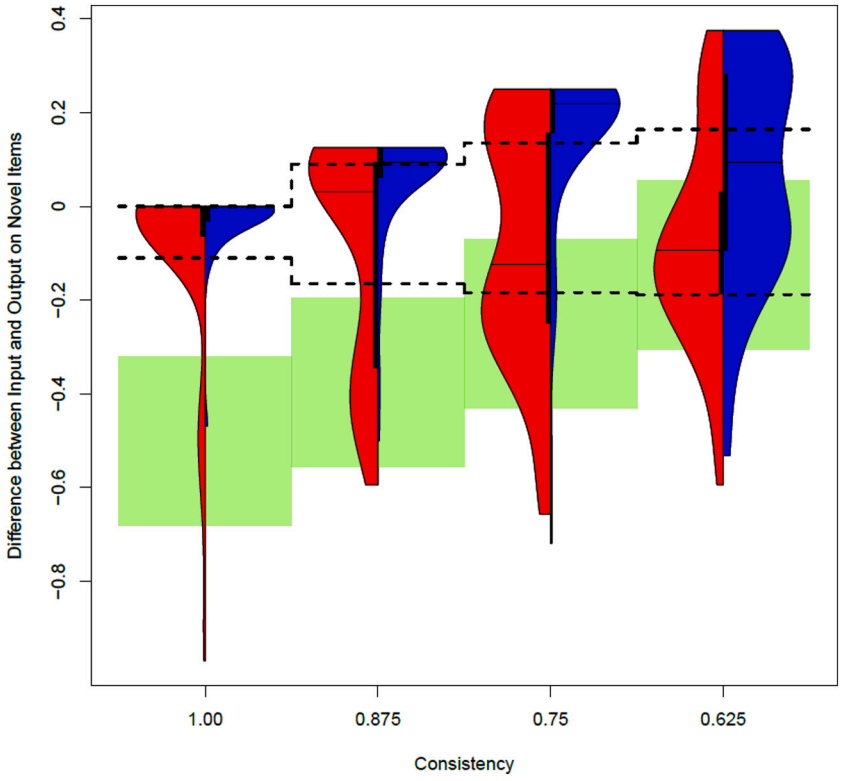

Fig. 9. Distributions of the difference between the proportion of consistent responses provided to novel items and the input proportion. Presentation consistency is on the x-axis. Singulative conditions are in red, Plural conditions in blue. Values at zero represent exact probability matching behavior, values above represent deviations towards majority regularization, and values below represent irregularization or minority regularization. Distributions are truncated at the ceiling for each condition. Dashed lines indicate the upper and lower bounds of the ninety-five percent (95%) binomial confidence interval for probability matching in each condition. The confidence interval is calculated including the uncertainty in estimating the underlying probability from only 32 trials. A green backdrop indicates the binomial confidence interval for guessing at random between two equally likely choices. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 11
Classification of individuals' output patterns for each inconsistent condition.

Condition	Minority Regularizer	Irregularizer	Probability Matcher	Majority Regularizer
Plural 0.625	6	2	35	36
Plural 0.75	2	3	12	62
Plural 0.875	0	5	23	51
Singulative 0.625	14	5	45	15
Singulative 0.75	5	26	24	24
Singulative 0.875	0	27	20	32
Total	27	68	159	220

Table 12
Classification of individuals' output patterns for each consistent condition.

Condition	Irregularizer	Probability Matcher
Plural 1.00	4	75
Singulative 1.00	12	67
Total	16	142

mechanism.

To further evaluate preservatory mechanisms as the basic mechanism for the results, we take this mechanism as a null hypothesis. We classify each particular participant as a probability matcher if their response pattern on novel items falls within the ninety-five percent (95%) confidence interval for the expected number of consistent responses for the input proportion. That is, a probability matcher is someone whose outputs are not significantly different from the input. We classify a participant as a majority regularizer if their response pattern falls above the high end of the interval. Minority regularizers are below the low end of the confidence interval, and, in addition, the proportion of the minority system in their output is greater than the proportion of the dominant system in the input. The remaining group of irregularizers have response patterns that fall in between the probability matchers and the minority regularizers. Table 11 shows how participants in the inconsistent conditions fell into each category.

Overall, far fewer than half of the participants (159 out of 474) fall within the confidence interval for probability matching. If the adult language learners in our experiment were, in general, probability matchers, we would expect 450 participants to fall within this interval. In every inconsistent condition, there is a population of participants for each possible output pattern, except for minority regularizers, which are not found in the relatively consistent 0.875 conditions. The differences between conditions are a matter of both degree and kind; sometimes the dominant output pattern is probability matching, and sometimes it is regularization, depending on the condition. For example, a majority regularized in the Plural 0.75 and Plural 0.875 conditions, but a plurality irregularized in the Singulative 0.75 condition. A related observation is that the number of dominant regularizers in inconsistent conditions also depends on the space between the input probability and the ceiling at 1.0.. The number of dominant regularizers in the Plural 0.75 condition is greater than the number of such regularizers in the Plural 0.875 condition because there is more room for systematizing behavior to express itself in the 0.75 conditions than the 0.875 conditions (there is more space between input and ceiling). Nevertheless, the large number of dominant regularizers in the Plural 0.75 and Plural 0.875 conditions demonstrates a systematizing tendency in the Plural conditions (Table 12).[6] Turning now to the consistent conditions, there were only two possible output patterns: irregularization and probability matching.

The Singulative 1.0 and the Plural 1.0 see a majority probability

[6] The question also arises of whether the discrepancies between the inconsistent and consistent conditions might be due to the differential use of strategies other than learning probabilities, such as trying hypotheses based on semantic subclasses of items. Indeed, in the debriefing questionnaire, 19 participants from inconsistent conditions did mention considering a semantic strategy. However, this number of participants is much too small to account for the discrepancies.

Table 13
Chi-square tests of the output classification of learners in all consistencies.

Consistency	χ^2	df	p
0.625	14.38	2	<0.001
0.75	39.57	2	<0.001
0.875	19.68	2	<0.001
1.00	3.408	1	0.065

matching, but this is presumably because probability matching is the same as regularization for the 1.0 conditions – it is not possible to produce a given system with proportion greater than 1.0, and probability matching in those conditions represents perfect systematicity. They are therefore ambiguous between probability matchers and regularizers, although in general they have maintained the level of input consistency in their output.

Given the systematicity of the Plural 0.75, 0.875, and 1.0 conditions, the Plural 0.625 stands out by its lower fraction of majority regularizers. While the Plural 0.75 and 0.875 have considerable mass above the input frequency with almost none below, far more Plural 0.625 participants fall below the input frequency. As a result, the Plural 0.625 and Singulative 0.625 distributions exhibit substantial overlap. This observation represents yet another way in which the 0.625 conditions fail to continue the patterns set by the other conditions. We will consider possible explanations in the discussion.

Significance levels for the differences in participant classification between the Singulative and Plural conditions can be assessed using χ^2 tests (Table 13). To avoid the problem of low cell counts with χ^2 tests, the minority regularizers were combined with irregularizers in the analysis.

Each of the comparisons is significant except for the 1.00 condition, indicating that the marking system factor is related to different distributions of participant response patterns in all inconsistent conditions. By collapsing minority regularization, irregularization, and probability matching, we can test specifically for different rates of majority regularization between the Plural and Singulative conditions. This difference is significant for the 0.625 conditions ($\chi^2 = 9.71$, df = 1, $p = 0.001$), for the 0.75 conditions ($\chi^2 = 41.86$, df = 1, p $\leq$0.001), and for the 0.875 conditions ($\chi^2 = 13.75$, df = 1, $p \leq 0.001$). Conflating majority regularization and probability matching, the differences are still significant for 0.625 ($\chi^2 = 4.47$, df = 1, $p = 0.035$), for 0.75 ($\chi^2 = 24.4$, df = 1, p $\leq$0.001), and for 0.875 ($\chi^2 = 17.28$, df = 1, p $\leq$0.001). This indicates that the number of irregularizers and minority regularizers contributes to the overall differences between the groups. In sum, for all inconsistent conditions, marking system was a significant factor in predicting the distribution of participant response patterns.

4. Discussion

The purpose of the experiment was to extend our understanding of how inconsistent systems are learned and generalized. It complements previous artificial language learning studies on the same topic by exploring a different linguistic domain, and by using a new experimental paradigm that enabled us to recruit large numbers of experimental participants on the web and to examine the progress of learning during the training phase. In analysing our results, we first asked whether they reflect a learning mechanism that is fundamentally preservatory or systematizing. We also asked whether the learning is biased or unbiased, and insofar as it is biased, we asked whether the bias is universal or varies across participants.

We can immediately discount several of the possibilities we discussed, by comparing the results to the predictions of the preservatory Estes-Bush-Mosteller and beta-binomial Bayesian models, on the one hand, and the predictions of the unbiased systematizing models, on the other. This comparison is made graphically in Fig. 10. Each panel depicts the distributions of responses for participants in each condition, for the generalization trials in the test phase. In the upper panel, they are overlaid on the space of the beta-binomial Bayesian model presented above. In the lower plot, they are overlaid on the predictions of an unbiased systematizing mechanism.

The preservatory mechanisms fail to capture the large number of regularizers in the Singulative 0.875 condition, as well as other Singulative regularizers. The outcome for the Singulative 1.00 condition is also problematic for the biased preservatory learning mechanisms. If this bias is universal, this condition should display the greatest deviation from the input frequency of all the conditions. If individuals vary in their bias towards the more familiar pattern, this condition is expected to display the most variability. But in fact participants in the Singulative 1.00 condition displayed the least variability, as most were near ceiling.

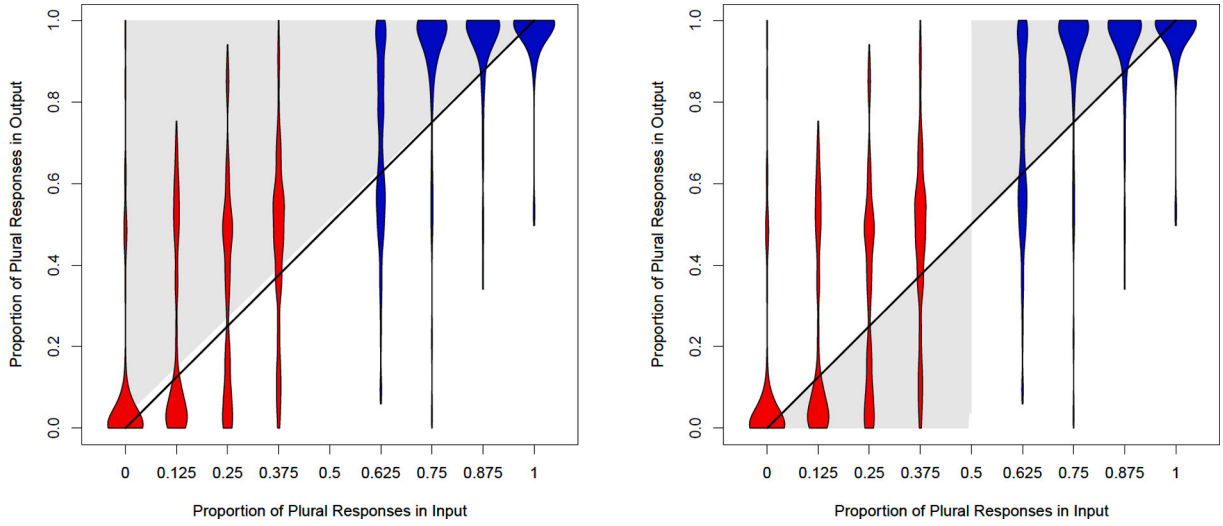

Fig. 10. Violin plots of the distributions in the data against the model predictions. Singulative conditions are in red, Plural conditions are in blue. The black lines show the input proportion, and the shaded area represents the area where the results could fall under any of the free parameter choices available in the model. The shaded area in the upper plot is for a preservatory beta-binomial Bayesian mechanism with the possibility of bias. The shaded area in the lower plot is for a systematizing mechanism without bias. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

These participants did not exhibit a bias towards a Plural system, instead preferring to retain the systematicity of the input.

Recall that over all conditions, the majority of participants produced responses outside of the ninety-five percent confidence interval for probability matching. The high rates of regularization in the inconsistent Plural conditions leave no doubt that participants in these conditions were systematizing the input. The pattern of variability in inconsistent Singulative conditions is inconsistent with the preservatory learning mechanisms. In those conditions, not only did many participants produce a response pattern inconsistent with probability matching, but they favored different patterns. Some Singulative participants favored regularization, but a substantial number irregularized. These observations are reflected in the regression analysis presented in Table 9. We conclude that our results are not explained by the preservatory models; specifically, these models cannot explain the variation in response patterns. It may be that no primarily preservatory mechanism can explain our findings. At least, our findings show that not all instances of regularization can be explained by a biased preservatory mechanism (a finding anticipated in Ferdinand (2015)). The actual mechanism must apparently be capable of producing primarily regularization along with response patterns that resemble probability matching in certain circumstances. But in such cases, it is still not preservatory in character.

The unbiased systematizing mechanism is more successful than the preservatory mechanism in capturing the behavior of the more consistent Singulative conditions. However, it does poorly with the inconsistent Singulative conditions, as indicated by the large number of Singulative participants whose outputs fall outside of the shaded area in the lower panel. Many participants irregularized, and some were minority regularizers. The Plural 0.625 condition also has a considerable number of participants outside of the shaded area. Under this mechanism, the dominant variant should be regularized regardless of how familiar it is – in this case, whether or not it is Plural. The significant main effect of marking system in Table 10 similarly cannot be explained with an unbiased systematizing mechanism, nor the effect of marking system in the participant classification documented in Table 11.

We are therefore led to the conclusion that an interaction of a substantive bias for plural-marking and a structural bias for systematization must be at work. The observed variability cannot be explained by the preservatory models, an unbiased systematizing mechanism (Fig. 10 right panel), or random guessing (Fig. 9) Instead, the results are understandable under the assumption that participants expect to see a Plural system (because it is the system they are more familiar with), and they also prefer a more consistent system. For the Plural conditions, the predictions associated with the two factors are aligned – more of the Plural system also results in more regular output. In the Singulative conditions, the two factors are in conflict, and the way they are resolved can lead either to irregularization or regularization.

In the next section, we will describe a mathematical mechanism that meets the desiderata. The model provides a family of functions that can capture behaviors ranging from exact probability matching to the highly nonlinear effects observed in our study. Before presenting the model, we wish to draw attention to a number of cautionary considerations regarding the relationship between the patterns observed in other studies and those that we observed. For one, our experiment provided learners with explicit feedback on each training trial. Some previous experiments also provided feedback before learners were tested for generalization behavior (Culbertson et al., 2012; Culbertson & Newport, 2015; Samara, Smith, Brown, & Wonnacott, 2017; Wonnacott et al., 2008), but these studies did not provide feedback on a trial-by-trial basis throughout training. Moreton and Pertsova (2016) compare rule-formation in an artificial language learning experiment for conditions in which learners do or do not receive feedback. While they report that the no-feedback condition yielded surprising high levels of rule-formation, differences between the conditions nonetheless exist. Thus, our training method may have favored rule-formation, which would in turn promote regularization, particularly in the more consistent conditions. The high level of regularization compared to previous studies that we observed therefore might be in part a product of the feedback mechanism. Secondly, our design focused on the morphological level and manipulated variability in marking system via homophony of an affix. Previous studies have examined probabilities such as probabilities of word co-occurrences or orderings. It is possible that learners handle inconsistencies arising through homophony differently from inconsistencies in word co-occurrences/orderings. More generally, imperfect learning of the assignment of stems to morphological classes is widely acknowledged to be a factor in historical change, particularly in analogical changes that tend to eliminate rare patterns (Daland & Sims, 2007; Lieberman, Jean-Baptiste, Jackson, Tang, & Nowak, 2007); this factor may be less pertinent for some other levels of representation. Thirdly, our training phase is relatively short. Although the behavior of most learners differed only moderately between the penultimate and final blocks of training (see section 3.7), it still remains possible that a much longer training phase would have yielded results more similar to prior studies. Finally in our experiment, the input consisted solely of examples of number marking, effectively highlighting the target contrast. In natural learning situations, examples of any given morphological contrast would be interspersed with other material. All of these factors potentially influence the learning process. Taken together, they may have induced greater deviations from probability-matching behavior than would be observed using other paradigms. Much further work would be needed to isolate the effects of these factors and to determine how they interact in shaping the overall relationship between input and output patterns.

5. Implications for mechanisms of learning

The key observation that a successful model must capture is the greater individual variation when the two biases are in conflict than when they are aligned. A fresh look at the outcomes for the Plural 0.625 and Singulative 0.625 conditions can yield further insights into what a formal model should look like. Recall that results from these two conditions were more similar than would be expected from the 0.75 and 0.875 conditions in times to completion in the training phase, correct responses on late blocks during training, and the distributions of responses to both seen and novel items in the test phrase. During the training phase, the familiar Plural 0.625 condition had little advantage in learnability over the less familiar Singulative 0.625 condition. During the test phrase, the amount of regularization in the Plural 0.625 was moderate in comparison to the Plural 0.75 and Plural 0.875. Although the 0.625 conditions have more room to the ceiling than the 0.75 conditions, and the difference in the ninety-five percent confidence intervals for probability matching is not great between the 0.75 and 0.625 conditions, the Plural 0.75 condition had over 1.7 times as many majority regularizers as the Plural 0.625. The Singulative 0.75 condition likewise had more majority regularizers than the Singulative 0.625.

Although initially perplexing, these results for the 0.625 conditions are anticipated in early experiments that evaluated probability matching as a mechanism for learning and making predictions. Edwards (1956, 1961) explored the ability of 120 basic airmen to predict whether the left or right space in a display would reveal a mark when it was uncovered. Feedback was provided on each of 1000 trials, and mark probabilities of 0.50, 0.60, and 0.70 were compared. Performance on the later blocks reveals the extent of learning from earlier blocks. For the 0.70 condition, participants in later blocks guessed the dominant location at rates of over 0.80. This deviation from probability matching towards a more optimal strategy is characterized by Edwards as “extreme-asymptote matching”, and it occurred to a lesser extent in the 0.60 condition. Myers and Atkinson (1964) review a variety of studies involving 250 or more training trials. In their own study manipulating payoff structures for three probabilities, 0.60, 0.70, and 0.80, they find stronger evidence of a systematizing choice mechanism for the 0.70 and 0.80 conditions than for the 0.60 condition. As they explain, Cotton and

Rechtschaffen (1958) also report stronger evidence of a systematizing choice mechanism for a 0.70 two-choice condition than a 0.60 two-choice condition. To summarize, regularizing behavior for consistency levels over approximately 0.70 has been reported for a large range of training lengths.

Results of this nature can be captured in a framework in which induction of a linguistic rule is a separate step from remembering examples. Mikheev (1997) develops such a framework in a natural language engineering context; Albright and Hayes (2002, 2003) modify and extend the approach in their analysis of psycholinguistic data. In Mikheev (1997), the strength of each affixation rule is determined from its observed rate of application in the set of words compatible with the rule: a discounting formula then uses the t-distribution to adjust the observed rate to reflect uncertainty based on the sample size. Specifically, the rule strength is adjusted to the lower limit of a specific confidence interval for the application rate, with Mikheev suggesting use of a two-tailed ninety percent (90%) confidence interval. Interestingly, for our 0.625 condition, this discounted value is 0.475, which means that the learner is not confident that the rule is more likely to apply than not. For the 0.75 condition, in contrast, the discounted value is 0.611, which is above 0.50.[7] We conjecture that a learner only adopts a rule if sufficiently confident that it is more likely to apply than not.

The theory of linguistic productivity developed in Yang (2005) also presupposes an explicit distinction between merely memorizing examples and forming a productive generalization (a "rule") based on the examples. Bringing to bear a number of assumptions about the mechanisms of lexical access, Yang argues that a generalization over N lexical examples is only efficient if the number of exceptions e is less than $N/\ln(N)$. This is Yang's "Tolerance Principle". Interestingly, for the 16 lemmas of our training set, the Tolerance Principle cutoff is 5 exceptions, which means that the 0.625 conditions would have too many exceptions to support generalization, but the 0.75 conditions do not.

An important difference between Mikheev's approach and Yang's approach is how they behave as the number of training examples increases. Mikheev predicts an increasing tolerance for the rate of exceptions (expressed as a probability) as the sample size increases. Yang predicts a decrease. However, the comparison of our study with the Edwards, Myers & Atkinson, and Cotton & Rechtschafen studies indicates that a consistency rate of about 0.60 results in weaker rule formation than consistencies of 0.70 or better, over a surprising range in the number of training examples. Clearly, this issue requires further research. Both approaches use universal thresholds to model aggregated data, thus offering no treatment of inter-participant variability. The Mikheev model is, however, readily extended by proposing variation in the free parameter in the discounting function. The Yang model has no free parameters, and would need to add one to achieve an explanation of the effects. Neither approach offers a formalization of how prior expectations might interact with variability in the propensity to systematize. Thus, we take away from these works a more general lesson: the importance of the distinction between memorizing examples and forming a generalization over them.

In a separate paper (Schumacher & Pierrehumbert, 2017) we develop, motivate, and validate a mathematical model, the Double Sigmoid Scaling (DSS) model, that uses two free parameters, b and c, to capture individual variation in the propensity to systematize and in the influence of prior expectations. The function that relates the input consistency to the output consistency is nonlinear. It has a double sigmoid shape with two inflection points and a flatter region in between them:

$$\frac{1}{1+e^{-\left(ln\left(\frac{p}{1-p}\right)+b\right)^{c}}} \tag{3}$$

The parameter c controls the degree of nonlinearity. The parameter b captures cognitive biases by warping the input-output relation towards the right or to the left on the [0,1] interval of probabilities. A rightward shift means that the individual requires a higher input proportion to regularize a pattern. A leftward shift means that the individual requires a lower input proportion to regularize a pattern. There is inter-participant variation in both free parameters. Inter-participant variability arises in this approach as mixtures of participants who are more or less prone to regularize, and who are differently influenced by their previous experience with the English Plural system. When the nonlinearity parameter is minimised ($c = 1$) and there is no cognitive bias ($b = 0$), the input-output relationship reduces to probability matching. In that case, the model reduces to a composition of the logit and logistic, which are inverses. Consequently, the input will be the same as the output.

We compared the DSS model to the beta-binomial model by fitting both models to the training data for each participant, and using the results to predict their performance on the novel test items. Fits were made using nls in R. The DSS achieves better overall fits, with a MSE of 0.032 as against 0.113 for the beta-binomial model. We also replicated an important characteristic of the DSS fits for the previous Schumacher and Pierrehumbert (2017) study. Because the participants were randomly assigned to experimental conditions after accepting the Amazon HIT, we expect the distributions of fitted parameter values to be highly similar across the different conditions. This expectation is much better met for the DSS than for the beta-binomial model; see Schumacher and Pierrehumbert (2017) for a more detailed discussion.

Only a few individuals are best described as having parameter settings with such low nonlinearity and low bias that their behavior approximates probability-matching. Most individuals exhibit moderate to extreme nonlinearity. The double-sigmoid shaped nonlinearity captures the finding that more participants regularized in the 0.75 conditions than in the 0.625 conditions. For most participants, the Singulative condition is associated with a rightward shift of the input-output function, compared to the Plural condition. This means that most people require less evidence to generalize a pattern when the pattern is similar to patterns they already know about. However, a few individuals display the opposite behavior, and the model can capture this outcome. As the consistency increases, there is sufficient statistical support for more and more individuals to regularize. As the consistency approaches 1.0, the output pattern becomes less and less sensitive to variation in both free parameters. As a consequence, the outputs in both the Plural 1.0 and the Singulative 1.0 conditions are expected to be clustered at ceiling (disregarding any noise that is not captured through variability in the two free parameters).

Because of its nonlinear input-output relationship, our model contrasts strongly with the approach of Hudson Kam and Newport (2005, 2009), which claims that linguistic learning in adults is fundamentally preservatory, with regularization sometimes observed because of sparse exposure to infrequent variants. Hudson Kam and Newport had fewer participants than the present study and they treat probability matching as the null hypothesis. One may therefore speculate that their conclusions differ from ours in part because of data aggregation or lower statistical power. It is also possible that linguistic differences and task differences between our study and the Hudson Kam and Newport studies may have promoted more regularizing behavior by our participants. Hudson Kam and Newport (2005) holds that linguistic learning in children is systematizing, a claim that is articulated further in Schuler et al. (2016), where Yang's Tolerance Principle is applied to analyze child data. The baseline data from the 20 adults in their study show a weaker systematizing tendency than the child data. Unfortunately due to low statistical power and aggregation of data, the extent of deviations

[7] Specific properties of the Albright & Hayes model have the result that both the 0.625 conditions and the 0.75 conditions in our study are expected to yield a slight degree of regularization. This model asymptotes to probability matching behavior as the size of the training set approaches infinity.

from probability match in the adult data cannot be determined. A strength of the model we have just laid out is that it can encompass slight or variable systematization within the same framework as extreme systematization.

Our proposal agrees with Reali and Griffiths' (2009) claim that language learning involves a prior expectation of systematicity. However, our findings are inconsistent with the specific explanation for regularity that those authors advance. In their preservatory Bayesian model, a weak prior combines with random sampling to produce regularization that is only visible in multi-generational chains. Their best fit model finds the parameter $\alpha < 0.1$, which is a negligible number in relation to the 32 distinct training trials in our study. Consequently, it cannot be responsible for the cases of extreme regularization observed in the experiment. This mechanism is also not consistent with the fact that the majority of our participants produced patterns outside of the ninety-five percent confidence interval for the underlying probability in their training set. Indeed, in Reali & Griffiths Fig. 2, the pooled data is closer to probability matching than the data for individual participants plotted in the second panel. This suggests that apparent probability matching in their study may have come about as a result of pooling data across participants, as was also noted by Ferdinand (2015).

Our results support the claim in Culbertson and Smolensky (2012) that substantive biases influence outcomes in language learning. Overall, participants in the Singulative conditions were less likely to systematize than participants in the Plural conditions. However, our results are not consistent with their particular implementation of substantive bias. As explained in connection with our discussion of the beta-binomial Bayesian model, their approach predicts a strong expression of Plural bias for the Singulative 1.00 condition, contrary to what we found. If the strength of the bias varies amongst individuals (implemented by varying the total count of applicable prior examples), it also incorrectly predicts that the Singulative 1.00 condition would display more variation than the inconsistent Singulative conditions. We conclude that substantive biases are not equivalent to remembered examples. Instead, our approach treats substantive biases at a more abstract level, modulating the threshold for forming a productive generalization. Substantive biases are more abstract in the sense that they do not reference specific properties or features of the input. Instead, they only affect how a learner behaves for a given amount of input.

Our results have implications for the theory of language variation and change. Language acquisition and language change are closely related, because language change occurs through iterated experience and production of linguistic forms. We began this paper by describing the tension between reports that the behavior of language learners is probability matching and patterns of language change documented in historical linguistics and sociolinguistics. If language learning were purely probability matching, the result would be long-term stable variation. More typically, however, highly variable patterns become regularized at historical time scales. While some instances of long-term apparently stable variation have been reported, Labov (2001) claims that such variation is socially conditioned. Although individual contexts show variability, the social associations of the different variants are not merely empirically present, they are also represented in speaker's and listener's minds, with the result that speakers modulate their output. While such cases display probabilistic variation, they are unlikely to be caused by a probability-matching learning mechanism. In such cases, the average production probabilities reflect aggregation over heterogeneous populations rather than learned probabilities over indexed variants. Baxter et al. (2009) carry out a detailed evaluation of a preservatory 'neutral evolution' model as applied to the evolution of New Zealand English. They show that even in such a model, individual speakers differ in their preferences for one variant over another, so that probability matching only arises through aggregation over speakers. The learning mechanism in our experiment proved to be more systematizing than preservatory. When participants were faced with an inconsistent system, most produced variable output patterns — but there were systematic discrepancies between the input and the output patterns. Inter-participant variability in generalization performance displayed far more cases of regularizing behavior than has been reported in many previous studies. It is possible that a confluence of methodological and linguistic choices in our study, particularly the use of ongoing feedback during training, may have promoted rule formation and therefore boosted the likelihood of regularizing behavior over what would occur in more naturalistic scenarios. However, tapping this extreme led us to develop a modelling framework in which any degree of systematizing can be captured, from none (yielding probability-matching behavior) to the absolute systematizing that would ensue from a threshold decision process. This framework provides tools for investigating the factors that determine the strength of people's propensity to systematize, potentially helping language scientists towards a more exact understanding of the mechanisms for regularization in different parts of the linguistic system.

The computational study of of New Zealand English by Baxter et al. (2009) ultimately concludes that the preservatory neutral evolution model is not realistic. The present study provides experimental support for their conclusion. Participants in our study differed greatly in the strength of their propensity to systematize. Such heterogeneity within the speech community may go towards explaining cases in which variability in a linguistic marker persists over many generations. The two competing variants may be produced predominantly by different speakers, with a subpopulation of speakers also reproducing this variation on an individual basis. There were important differences amongst individuals in their response to the innovative Singulative pattern. Participants who responded to the inconsistent system by systematizing the Singulative pattern can be interpreted as individuals with a strong propensity to systematize, but a weak commitment to the Plural. These results are consistent with computational simulations suggesting that the spread of innovative variants depends on a heterogeneous speech community including a critical mass of people who will adopt a variant even it is represents a minority pattern in the input (Blythe & Croft, 2012; Pierrehumbert et al., 2014).

6. Conclusion

We investigated the interaction of marking system and consistency in the acquisition of a morphological system through the use of number marking systems. We compared two number marking distinctions encoded by a single affix, a Plural system and a Singulative system. During the training phase, these were shown either consistently (1.00 of the items used the same system) or inconsistently (0.875, 0.75, or 0.625 of the items used one of the systems, with the remaining items using the other). In the test phase, participants generally regularized the dominant system if it was also familiar. They were also essentially at ceiling for the less familiar input pattern, provided that it was fully consistent. The results for inconsistent, less familiar systems were very different. Analyzed in detail, the results indicate that the language learning mechanism responsible for the results is fundamentally systematizing. However, individuals differ in the strength of their propensity to systematize. The results also reflect a bias towards the expected Plural pattern. However, we argued that the strength of this bias is not equivalent to some number of previously experienced examples, as in the beta-binomial Bayesian model. Instead it applies at an abstract level, modulating the level of statistical evidence that a participant requires to form a productive generalization. That is, a potential new generalization is either facilitated or inhibited on the basis of its similarity to a previously known generalization. Thus, the results challenge theories in which regularization arises indirectly from a single free parameter in a preservatory learning mechanism. Instead, they favour a theory with at least two free parameters, which capture a participant's propensity to systematize and their bias; the DSS model that we sketch provides an concrete instantiation of such a theory. The results also tend to support theories of language change in which heterogeneity in the speech community plays a key role.

CRediT authorship contribution statement

R. Alexander Schumacher: Conceptualization, Methodology, Visualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Janet B Pierrehumbert:** Conceptualization, Methodology, Investigation, Resources, Supervision, Funding acquisition, Writing - original draft, Writing - review & editing.

Acknowledgments

We would like to thank Jen Hay and Cynthia Clopper for comments, and statisticians Patrick LaShell and Shane Pederson for consultation on the statistical analysis. This project was made possible through a grant to Northwestern University from the John Templeton Foundation (Award ID 36617). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Appendix A. Words and images

baref	barefyl	neiat	neiatyl
batif	batifyl	nuoid	nuoidyl
cufig	cufigyl	panig	panigyl
demil	demilyl	priud	priudyl
ewuar	ewuaryl	raeif	raeifyl
feabr	feabryl	reoud	reoudyl
gomur	gomuryl	sauin	sauinyl
guliw	guliwyl	sloaq	sloaqyl
isoer	isoeryl	slour	slouryl
lopas	lopasyl	suagr	suagryl
mieaf	mieafyl	sugin	suginyl
nehad	nehadyl	touaf	touafyl
winak	winakyl	vonuf	vonufyl
yimag	yimagyl	wasok	wasokyl
yosic	yosicyl	wiben	wibenyl
yuseg	yusegyl	wifud	wifudyl

Training Phase Image	Label	Test Phase Image	Label
	banana		asteroid
	basket		candle
	bear		bok choy
	dress		berry
	fish		gate
	flower		goblet
	fox		hat

(*continued on next page*)

(*continued*)

Training Phase Image	Label	Test Phase Image	Label
			
	human		gourd
	cashewnut		wrench
	chair		asterisk
	pitcher		pet
	squid		mushroom
	turnip		cookie
	door		space pod
	blanket		sheep
	duck		bug

Appendix B. Sample blocks

Singulative 100					
Image name	Word 1	Word 2	Number	Item type	Suffixed Form Correct?
fish	raeifyl	raeif	five	dominant	FALSE
basket	sauin	sauinyl	one	dominant	TRUE
fish	raeifyl	raeif	one	dominant	TRUE
flower	demil	demilyl	five	dominant	FALSE
chair	yosicyl	yosic	one	dominant	TRUE
squid	mieafyl	mieaf	five	dominant	FALSE
blanket	feabr	feabryl	five	dominant	FALSE
door	sugin	suginyl	one	dominant	TRUE

(*continued*)

Singulative 100

Singulative 87.5

Image name	Word 1	Word 2	Number	Item type	Suffixed Form Correct?
fish	reoud	reoudyl	one	dominant	TRUE
dress	guliw	guliwyl	five	dominant	FALSE
cashewnut	ewuar	ewuaryl	one	dominant	TRUE
fish	reoudyl	reoud	five	dominant	FALSE
cashewnut	ewuaryl	ewuar	five	dominant	FALSE
fox	neiat	neiatyl	one	dominant	TRUE
dress	guliw	guliwyl	one	dominant	TRUE
basket	nehadyl	nehad	one	minority	FALSE

Singulative 75

Image name	Word 1	Word 2	Number	Item type	Suffixed Form Correct?
fox	winakyl	winak	one	dominant	TRUE
fox	winakyl	winak	five	dominant	FALSE
cashewnut	nuoidyl	nuoid	five	dominant	FALSE
chair	baref	barefyl	five	minority	TRUE
flower	suagryl	suagr	one	dominant	TRUE
banana	neiatyl	neiat	one	dominant	TRUE
pitcher	isoer	isoeryl	five	dominant	FALSE
chair	barefyl	baref	one	minority	FALSE

Singulative 62.5

Image name	Word 1	Word 2	Number	Item type	Suffixed Form Correct?
squid	touaf	touafyl	one	minority	FALSE
fox	cufig	cufigyl	five	dominant	FALSE
dress	sauin	sauinyl	five	dominant	FALSE
fox	cufigyl	cufig	one	dominant	TRUE
squid	touaf	touafyl	five	minority	TRUE
door	yosicyl	yosic	five	dominant	FALSE
door	yosicyl	yosic	one	dominant	TRUE
cashewnut	suagr	suagryl	five	minority	TRUE

Plural 100

Image name	Word 1	Word 2	Number	Item type	Suffixed Form Correct?
fox	neiat	neiatyl	one	dominant	FALSE
fish	sloaqyl	sloaq	five	dominant	TRUE
blanket	suginyl	sugin	five	dominant	TRUE
door	priud	priudyl	one	dominant	FALSE
dress	demilyl	demil	one	dominant	FALSE
duck	suagryl	suagr	five	dominant	TRUE
squid	batifyl	batif	one	dominant	FALSE
fox	neiat	neiatyl	five	dominant	TRUE

Plural 87.5

Image name	Word 1	Word 2	Number	Item type	Suffixed Form Correct?
fox	ewuar	ewuaryl	five	dominant	TRUE
bear	sauinyl	sauin	five	dominant	TRUE
fox	ewuar	ewuaryl	one	dominant	FALSE
basket	neiat	neiatyl	one	minority	TRUE
fish	yuseg	yusegyl	one	dominant	FALSE
squid	suagryl	suagr	five	dominant	TRUE
bear	sauinyl	sauin	one	dominant	FALSE
squid	suagryl	suagr	one	dominant	FALSE

Plural 75

Image name	Word 1	Word 2	Number	Item type	Suffixed Form Correct?
fox	vonufyl	vonuf	one	dominant	FALSE
pitcher	winakyl	winak	one	dominant	FALSE
turnip	lopasyl	lopas	one	minority	TRUE
pitcher	winak	winakyl	five	dominant	TRUE
fish	sugin	suginyl	one	dominant	FALSE
bear	suagr	suagryl	five	dominant	TRUE
turnip	lopasyl	lopas	five	minority	FALSE
dress	isoer	isoeryl	five	dominant	TRUE

Plural 62.5

Image name	Word 1	Word 2	Number	Item type	Suffixed Form Correct?
cashewnut	reoud	reoudyl	one	minority	TRUE

(*continued on next page*)

(*continued*)

Plural 62.5					
pitcher	nuoid	nuoidyl	one	dominant	FALSE
door	nehad	nehadyl	five	dominant	TRUE
turnip	raeifyl	raeif	one	dominant	FALSE
flower	barefyl	baref	five	minority	FALSE
turnip	raeifyl	raeif	five	dominant	TRUE
door	nehad	nehadyl	one	dominant	FALSE
flower	baref	barefyl	one	minority	TRUE

Appendix C. Model specifications

C.1. Training phase

glmer(correct_answer ~ familiarity * consistency + trial_number + lag_1_trial + (1|image) + (1 + trial_number | participant_id), data = training_phase, family = binomial(link = "logit"), control = glmerControl(optimizer = "bobyqa")).

C.2. Test phase (seen)

glmer(consistent_response ~ familiarity * consistency + (1|image) + (1 | participant_id), data = test_phase_seen, family = binomial(link = "logit"), control = glmerControl(optimizer = "bobyqa")).

C.3. Test phase (novel)

glmer(consistent_response ~ familiarity * consistency + (1|image) + (1 | participant_id), data = test_phase_novel, family = binomial(link = "logit"), control = glmerControl(optimizer = "bobyqa")).

References

Albright, A., & Hayes, B. (2002). Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on morphological and phonological learning-volume 6* (pp. 58–69). Association for Computational Linguistics.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition, 90*(2), 119–161.

Anderson, S. R. (1985). Inflectional morphology. In T. Shopen (Ed.), *Language typology and syntactic fieldwork vol. III* (pp. 150–201). Cambridge: Cambridge University Press.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*(3), 372–400.

Baayen, H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2).

Baer-Henney, D., Kügler, F., & van de Vijver, R. (2014). The interaction of language-specific and universal factors during the acquisition of morphophonemic alternations with exceptions. *Cognitive Science, 39*(7), 1537–1569.

Baerman, M. (2007). Morphological reversals. *Journal of Linguistics, 43*(1), 33–61.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–193). New York: Routledge.

Baxter, G. J., Blythe, R. A., Croft, W., & McKane, A. J. (2009). Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change, 21*(2), 257–296.

Blythe, R. A., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language, 88*(2), 269–304.

de Boer, B. (2001). *The origins of vowel systems: Studies in the evolution of language*. Oxford: Oxford University Press.

de Boer, B., & Zuidema, W. (2010). Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior, 18*(2), 141–154.

Brooks, P. J., Braine, M. D., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language, 32*(1), 76.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review, 58*(6), 413.

Bybee, J., & Beckner, C. (2010). Usage-based theory. In B. Heine, & H. Narrog (Eds.), *The Oxford handbook of linguistic analysis* (pp. 827–856). Oxford: Oxford University Press.

Bybee, J. L., & Slobin, D. I. (1982). Rules and schemas in the development and use of the English past tense. *Language*, 265–289.

Chen, G., Gully, S. M., Whiteman, J.-A., & Kilcullen, R. N. (2000). Examination of relationships among trait-like individual differences, state-like individual differences, and learning performance. *Journal of Applied Psychology, 85*(6), 835–847.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language, 30*(3), 637–669.

Cotton, J. W., & Rechtschaffen, A. (1958). Replication reports: Two and three choice verbal conditioning phenomena. *Journal of Experimental Psychology, 56*(1), 96.

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

Culbertson, J., & Smolensky, P. (2012). A Bayesian model of biases in artificial language learning: The case of a word order universal. *Cognitive Science, 36*(8), 1468–1498.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition, 122*(3), 306–329.

Daland, R., Sims, A., & Pierrehumbert, J. (2007). Much ado about nothing: a social network model of Russian paradigmatic gaps. In *Proceedings of the 45th annual meeting of the association for computational linguistics. Prague, Czech Republic, June 24th–29th, 2007*.

Dimmendaal, G. J. (2000). Number marking and noun categorization in Nilo-Saharan languages. *Anthropological Linguistics, 42*(2), 214–261.

Edwards, W. (1956). Reward probability, amount, and information as determiners of sequential two-alternative decisions. *Journal of Experimental Psychology, 52*(3), 177.

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology, 62*(4), 385.

Ellis, N. C. (2015). Implicit and explicit learning: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*. Amsterdam: John Benjamins.

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N. (2001). Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. https://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en.

Estes, W. K. (1957). Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika, 22*(2), 113–132.

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua, 120*(8), 2061–2079.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences, 109*(44), 17897–17902.

Ferdinand, V. A. (2015). *Inductive evolution: Cognition, culture, and regularity in language*. Ph.D Dissertation. University of Edinburgh.

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

Ferdinand, V. A., Thompson, B., Kirby, S., & Smith, K. (2013). Regularization in a non-linguistic domain. *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*(35).

Gallistel, C. R. (1990). *The Organization of Learning (learning, development, and conceptual change)*. Cambridge: MIT Press.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America, 101*(36), 13124–13131.

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not weird: The promise of the internet in reaching more diverse samples. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X10000300.

Grimm, S. (2012a). Individuation and inverse number marking in Dagaare. In *Count and mass across languages* (pp. 75–98). Oxford University Press.

Grimm, S. (2012b). *Number and individuation*. Doctoral dissertation: Stanford University.
Haspelmath, M., & Karjus, A. (2017). Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency. *Linguistics, 55*(6), 1213–1235.
Hayes, B., Siptar, P., Zuraw, K., & Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language, 85*(4), 822–863.
Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1264–1284.
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.
Hulstijn, J. H. (2015). Explaining phenomena of first and second language acquisition with the constructs of implicit and explicit learning. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*. Amsterdam: John Benjamins.
Jackendoff, R. (1975). Morphological and semantic regularities in the lexicon. *Language, 51*(3), 639–671.
Janse, E., & Newman, R. S. (2013). Identifying nonwords: Effects of lexical neighborhoods, phonotactic probability, and listener characteristics. *Language and Speech, 56*(4), 421–441.
Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General, 133*(3), 450–467.
Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *Evolutionary Computation, IEEE Transactions on, 5*(2), 102–110.
Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, 105*(31), 10681–10686.
Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences, 104*(12), 5241–5245.
Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change, 1*(3), 199–244.
Kurumada, C., & Grimm, S. (2019). Predictability of meaning in grammatical encoding: Optional plural marking. *Cognition, 191*.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in linear mixed effects models* (CRAN).
Labov, W. (1994). *Principles of linguistic change, Vol. I: Internal Factors*. Oxford: Blackwell.
Labov, W. (2001). *Principles of linguistic change, Vol. II: Social Factors. Oxford: Blackwell*.
Lakoff, G. (1971). *Syntactic irregularity*. New York: Holt, Rinehart & Winston.
Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63*(8), 1279–1292.
Lieberman, E., Jean-Baptiste, M., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature, 449*, 713–716.
Ma, T., & Komarova, N. L. (2017). Mathematical modeling of learning from an inconsistent source: A nonlinear approach. *Bulletin of Mathematical Psychology, 79*, 635–661.
Mandelshtam, Y., & Komarova, N. L. (2014). When learners surpass their models: Mathematical modeling of learning from an inconsistent source. *Bulletin of Mathematical Biology, 76*(9), 2198–2216.
Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition, 46*(1), 53–85.
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.
Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics, 23*(2), 405–423.
Moreton, E., & Pertsova, K. (2016). Implicit and explicit processes in Phonotactic learning. In *Proceedings of the 40th annual boston university conference on language development, Boston, United States*.
Myers, J. L., & Atkinson, R. C. (1964). Choice behavior and reward structure. *Journal of Mathematical Psychology, 1*(1), 170–203.
Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua, 108*(2), 95–117.
Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge: MIT Press.
Paul, E. J., & Ashby, F. G. (2013). A neurocomputational account of how explicit learning bootstraps early procedural learning. *Frontiers in Computational Neuroscience, 18*. https://doi.org/10.3389/fncom.2013.00177.
Perfors, A. (2012). Probability matching vs over-regularization in language: Participant behavior depends on their interpretation of the task. In *Proceedings of the 34th annual meeting of the cognitive science society*.
Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development, 12*(2), 138–155.
Pierrehumbert, J. B. (2012). The dynamic lexicon. In A. Cohn, M. Huffman, & C. Fougeron (Eds.), *Handbook of laboratory phonology* (pp. 173–183). Oxford University Press.
Pierrehumbert, J. B., Stonedahl, F., & Daland, R. (2014). *A model of grassroots changes in linguistic systems. arXiv:1408.1985v1*.
Plunkett, K., & Marchman, V. (1992). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*(1), 21–69.
Rácz, P., Beckner, C., Hay, J. B., & Pierrehumbert, J. B. (2020). Morphological convergence as on-line lexical analogy. *Language*. in press.
Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language|. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2017.00051.
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. (2014). *Mathematical modeling of language learning*. Ph.D Dissertation: University of California, Irvine.
Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned variation. *Cognitive Psychology, 94*, 85–114.
Sankoff, G., & Blondeau, H. (2007). Language change across the lifespan: /r/ in Montreal French. *Language, 83*(3), 560–588.
Saxton, M. (2000). Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language, 20*(60), 221–252.
Saxton, M., Backley, P., & Gallaway, C. (2005). Negative input for grammatical errors: Effects after a lag of 12 weeks. *Journal of Child Language, 32*(3), 643–672.
Schmidt, R. (2012). Attention, awareness, and individual differences in language learning. In W. M. Chan, K. N. Chin, S. K. Bhatt, & I. Walker (Eds.), *Perspectives on individual characteristics and foreign language education. Berlin, Boston: De Gruyter Mouton* (pp. 27–50).
Schuler, K., Yang, C., & Newport, E. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. In *Proceedings of the 38th annual conference of the cognitive science society. cognitive science society*.
Schumacher, & Pierrehumbert. (2017). Prior expectations in linguistic learning: A stochastic model of individual differences. In *Proceedings of the 38th annual meeting of the cognitive science society*. Cognitive Science Society.
Schumacher, R. A., Pierrehumbert, J. B., & LaShell, P. (2014). Reconciling inconsistency in encoded morphological distinctions in an artificial language. In *Proceedings of the 36th annual conference of the cognitive science society. cognitive science society*.
Scobbie, J. (2006). Flexibility in the face of incompatible VOT systems. In L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *8. Laboratory phonology* (pp. 367–392). Mouton de Gruyter.
Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198–213.
Skehan, P. (1998). *A cognitive approach to language learning. Oxford applied linguistics*. Oxford: Oxford University Press.
Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition, 116*(3), 444–449.
Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263). Association for Computational Linguistics.
Stuart-Smith, J., Pryce, G., Timmins, C., & Gunter, B. (2013). Television can also be a factor in language change: Evidence from an urban dialect. *Language, 89*(3), 501–536.
Thomason, S. G., & Kaufman, T. (1988). *Language contact, creolization, and genetic linguistics*. University of California Press. https://doi.org/10.3389/fpsyg.2014.00634.
van de Vijver, R., & Baer-Henney, D. (2012). Sonority intuitions are provided by the lexicon. In S. Parker (Ed.), *The sonority controversy* (pp. 195–218). De Gruyter Mouton: Berlin, Boston.
van de Vijver, R., & Baer-Henney, D. (2014). Developing biases. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00634.
Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107*(2), 729–742.
Wagner, S. E., & Sankoff, G. (2011). Age grading in the Montréal French inflected future. *Language Variation and Change, 23*(3), 275–313.
Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.
Wonnacott, E., & Newport, E. L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In *Proceedings of the 29th annual boston university conference on language development* (pp. 663–673).
Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology, 56*(3), 165–209.
Yang, C. (2005). On productivity. In *Linguistic variation yearbook 5* (pp. 265–302). John Benjamins Publishing Company.