Acquiring variation in an artificial language: children and adults are sensitive to socially conditioned linguistic variation

Anna Samara[1], Kenny Smith[2], Helen Brown[3], & Elizabeth Wonnacott[1]

[1]*Division of Psychology and Language Sciences, University College London*

[2] *School of Philosophy, Psychology and Language Sciences, University of Edinburgh*

[3]*Department of Psychology, University of Warwick*

Contact details:

Anna Samara

Division of Psychology and Language Sciences

University College London

Chandler House, 2 Wakefield Street

London, WC1N 1PF

United Kingdom

Email: anna.samara@ucl.ac.uk

Abstract

Languages exhibit sociolinguistic variation, such that adult native speakers condition the usage of linguistic variants on social context, gender, and ethnicity, among other cues. While the existence of this kind of socially conditioned variation is well-established, less is known about how it is acquired. Studies of naturalistic language use by children provide various examples where children's production of sociolinguistic variants appears to be conditioned on similar factors to adults' production, but it is difficult to determine whether this reflects knowledge of sociolinguistic conditioning or systematic differences in the input to children from different social groups. Furthermore, artificial language learning experiments have shown that children have a tendency to eliminate variation, a process which could potentially work against their acquisition of sociolinguistic variation. The current study used a semi-artificial language learning paradigm to investigate learning of the sociolinguistic cue of speaker identity in 6-year-olds and adults. Participants were trained and tested on an artificial language where nouns were obligatorily followed by one of two meaningless particles and were produced by one of two speakers (one male, one female). Particle usage was conditioned deterministically on speaker identity (Experiment 1), probabilistically (Experiment 2), or not at all (Experiment 3). Participants were given tests of production and comprehension. In Experiments 1 and 2, both children and adults successfully acquired the speaker identity cue, although the effect was stronger for adults and in Experiment 1. In addition, in all three experiments, there was evidence of regularization in participants' productions, although the type of regularization differed with age: children showed regularization by boosting the frequency of one particle at the expense of the other, while adults regularized by conditioning particle usage on lexical items. Overall, results demonstrate that children and adults are sensitive to speaker identity cues, an ability which is fundamental to tracking sociolinguistic variation, and that children's well-established

tendency to regularize does not prevent them from learning sociolinguistically conditioned variation.

Keywords: Artificial Language Learning; Language Acquisition; Statistical Learning; Sociolinguistic Variation; Regularization.

1. Introduction

Variation is ubiquitous in natural language and occurs at all levels of analysis, be it phonetic, morphological, syntactic, semantic, or lexical. However, the usage of linguistic variants tends to be *conditioned*, so that variation is rarely, possibly never, fully unpredictable. A clear example of deterministic conditioning is the regular past tense marker in English (written as –ed), which can be realized as [t], [d], or syllabic [ɪd] (as in *liked*, *loved*, *hated*), with the choice of variant dependent on the phonological features of the final segment of the stem. However, conditioning contexts may also be probabilistic. For example, in English, a final (–t, d) segment in a final cluster is variably deleted (*nest* vs. *nes'*) with the probability of deletion affected by a variety of linguistic factors, including the phonological features of the following segment (e.g., /t/ or /d/ followed by an obstruent is more likely to be deleted than /t/ or /d/ followed by a liquid), morphological class (e.g., final /t/ or /d/ in monomorphemes delete more frequently than weak past tense forms) and the presence of a following pause, as well as social factors such as the speaker's gender and social context. We will refer to patterns of variation involving these latter kinds of social conditioning factors as *sociolinguistic*.

Although the existence of sociolinguistic variation is well-established in adult language, less is known about how children respond to socially conditioned variation during acquisition. For example, can young children pick up on the fact that different speakers use different variants, even when that relationship is probabilistic (as is often the case in adult languages)? How does the learning of socially conditioned variation interact with their well-demonstrated tendency to regularize experimenter-created miniature languages that exhibit fully unpredictable variation? The current paper explores these questions using a statistical learning framework, comparing children's and adults' learning of a form of sociolinguistic variation (variable forms cued by speaker identity) in an artificial language.

1.1. Sociolinguistic variation in adult learners

In his seminal work, Labov (1963) demonstrated that linguistic variation may be subject to probabilistic, extra-linguistic social constraints. He found that the pronunciation of some vowels on Martha's Vineyard (an island off the coast of Massachusetts) was changing from the standard American pronunciation, and that this was affected by social identity: age, occupation, and social group were all important factors, with those who identified most closely as natives of the island using the incoming forms most frequently. Social influences on linguistic choice were further established in Labov's (1966) "New York department store study", which examined differences in the pronunciation of postvocalic /r/ (e.g., as occurring in *fourth*, *floor*) in New York speech. Rhotic varieties (where /r/ is pronounced post-vocalically) were associated with prestige in New York, and were favoured by employees of an upper-middle class shopping center, whereas /r/-dropping was more frequently used by employees in a shopping center with a more working-class clientele. This work also revealed *within-speaker* variation and correlations between language style and context, with rhoticity notably higher in the context of careful or emphatic speech, particularly so among employees catering for a middle-class clientele.

Over the last 50 years, a wide range of sociolinguistic studies have since confirmed Labov's early findings, showing that in adult language, production of phonological and grammatical variation is in accordance with a number of extra-linguistic social factors. For example, there is a large literature demonstrating differences in male and female language use (e.g., Labov, 1966, 2001; Neu, 1980; Trudgill, 1974; Wolfram, 1969). This shows that speakers may associate certain variants with gender and avoid variants they perceive as gender-inappropriate (Eckert & McConnell-Ginet, 1999; Cameron, 2005). Women may also use forms that represent stable social variables (i.e., "standard" forms) in their speech more

frequently than men (Labov, 1990). Age (e.g., speech before and after adolescence) and ethnicity (i.e., ethnically marked ways of speaking) have been further shown to exert an influence on linguistic choices (e.g., Eckert, 1997; Foulkes, Docherty, & Watt, 2005; Hoffman & Walker, 2010; Knack, 1991; Mendoza-Denton, 2008).

1.2. Sociolinguistic variation in child language

Traditionally, sociolinguistic research has focused on adult language users rather than addressing questions of acquisition. An early reason was an assumption (originating with Labov, 1964) that child speech is monostylistic early in development, with little sociolinguistic competence before the age of 12. The methodological complications in working with children (e.g., the inadequacy of the sociolinguistic interview for the elicitation of speech in young children, practical difficulties in obtaining corpus of data large enough to locate sociolingustic variation; Foulkes, Docherty, & Watt, 2001; Roberts, 1994) also proved a barrier. Labov (1989) was among the first to empirically address the question of *when* children first exhibit adult-like command of sociolinguistic variables. He analysed the speech of three Philadelphian children (aged 6, 7, and 9 years) during family interaction, focusing on two sociolinguistic target variables, namely (–t, d) deletion and alternations between /ɪŋ/ and /ɪn/ in progressive inflection on verbs. In both of the 6- and 7-year-old's speech, there was some evidence that some patterns of variation seen in their parents' speech and their local dialect were present (for example, /t/ and /d/ were hardly ever deleted before pauses), although other constraints were not mastered (for example, unlike adults, they did not show greater likelihood of deletion in certain phonological contexts). While other work has not replicated these precise findings with this variable, the emerging literature confirms that conditioned variation *does* appear in children's speech from early ages, and that the older the children get, the more factors condition variation in their speech (Shin, 2016). Importantly,

the exact age of mastery differs across variables and constraints, possibly due to their differential distributional tendencies in the input and their complexity (Shin, 2016). For example, Roberts (1997) studied (-t,-d) deletion in sixteen 3-4 year olds in Philadelphia using an elicitation paradigm. She found that even these young children had acquired some relevant grammatical and phonological constraints (in contrast to Labov's findings); however, patterns of (t- d) deletion did not differ as a function of formal versus informal speech (elicited using different activities), or style of interaction (studied by eliciting speech addressed towards different interlocutors such as the experimenter or a puppet). A similar pattern was found by Smith, Durham, and Fortune (2009) in data from 11 Scottish 2- to 3-year-olds in naturalistic interactions with their caregivers: children largely mirrored the patterns of phonological conditioning present in their caregivers' speech, but did not have the same patterns of conditioned variation for morphological class and style shifting. Smith, Durham, and Richards (2013; see also Smith, Durham, & Fortune, 2007) used similar methods with 29 2- to 4-year-olds and analysed a number of lexical, phonological and morpho-syntactic variables. For these variables, they found good evidence that linguistic constraints had been acquired, even in the youngest children, who used these variables similarly to their caregivers. Style-shifting constraints (different usage of the variables in teaching vs. play and routine vs. discipline), were only found in older children (> 3 years) and only for certain variables. Interestingly, caregiver's speech during these interactions did not exhibit evidence of style shifting for many variables, highlighting the importance of considering the actual input available to children (see also Foulkes et al., 2005), as opposed to the norms of adult-adult speech.

Research has also explored whether children are sensitive to sociolinguistic conditioning based on various components of speaker identity. For example, Díaz-Campos (2005) looked at usage of phonological variation in Spanish (retention or deletion of

intervocalic /d/) and found that, like adults, 5-year-old children from higher socioeconomic backgrounds showed more usage of the standard form (see also Reid, 1978, Romaine, 1978, 1984). Note, however, that this is likely directly tied to the frequency of the standard form in their input. Although the children did show some style shifting (indicating some knowledge of the differing social values of the variants), there was no evidence that they were actually aware of the differential usage by different groups of speakers.

Gender differences have also been explored. For example, an early study by Romaine (1978) looked at socially marked variants of word-final /r/ in Scottish English and found that, as with adults, girls used the standard variants more frequently. Habib (in press) studied usage of two socially marked phonological variants of /q/ in Arabic, rural [q] versus urban [ʔ], and found that, like adults, children showed gender differences, with girls using the newer, more urban variant more frequently than boys (although not until after the age of 8). However, it is again possible that gendered usage of linguistic variants could stem from input differences. There is evidence that such differences can occur: Foulkes et al. (2005) analyzed parents' use of /t/-glottaling in contexts with clear sociolinguistic patterning for adults and found that speech to girls generally contained more standard variants than speech to boys. This work highlights the difficulty of establishing when and whether children's language reflects internal knowledge of sociolinguistic conditioning contexts, as opposed to simply reflecting frequencies in their input.

In sum, there is good evidence in the literature that children exhibit use of linguistic variants from an early age and that their usage of these variants may be constrained by some of the probabilistic conditioning factors which they encounter in their input. In particular, it is clear that linguistic constraints (phonological, morphological, lexical) can be learned in early childhood, although these may emerge at different ages for different variants. There is also evidence that, in some cases, children's usage of linguistic variants may be consistent with

adult social norms. However, on the basis of naturalistic data it can be difficult to establish whether and when they have actually acquired the relevant sociolinguistic conditioning cues, or whether their output merely reflects systematic biases in the input to children from different social groups. In the current work, we explore children's learning of sociolinguistically conditioned variation using the *artificial language learning* methodology, which allows direct experimental control over the input to be learnt.

1.3. Artificial language learning and the acquisition of variation

Artificial language learning–where learners are exposed to and tested on experimenter-created miniature languages–provides a methodology for exploring the types of linguistic regularities which learners can acquire. There is evidence that artificial languages can be processed similarly to natural languages (Ettlinger, Morgan-Short, Faretta-Stutenberg, & Wong, 2015; Magnuson, Tanenhaus, Aslin, & Dahan, 2003; Wonnacott, Newport, & Tanenhaus, 2008). This method now has a long history in the field of language acquisition (e.g., Braine, 1963) and has been used to address such topics as word segmentation (e.g., Saffran, Aslin, & Newport, 1996), word learning (e.g., Smith & Yu, 2008; Yu & Smith, 2007), the acquisition of phonology (Chambers, Onishi, & Fisher, 2010) and syntax (Frigo & McDonald, 1998; Gerken, Wilson, & Lewis, 2005; Reeder, Newport, & Aslin, 2013; Wonnacott et al., 2008, Wonnacott, Boyd, Thomson, & Goldberg, 2012) in both adults and children. Much of this work focuses on how sensitivity to statistically based patterns in the input can aid the extraction of linguistic generalizations.

Most relevant to the current study is a set of statistical learning experiments looking at the learning of artificial languages which exhibit *unpredictable variation.* In a seminal study, Hudson Kam and Newport (2005) exposed adults and children (5- & 6-years old) to an

artificial language in which nouns could be followed by particles[1]. These particles occurred variably in the input presented to the learners and could be included or omitted: this variation was probabilistic in that one form had higher frequency in the input than the other (e.g., in one condition, particles were included 60% of the time a noun was produced and omitted 40% of the time) but, importantly, unlike natural languages, there were no further conditioning cues (either deterministic or probabilistic) which determined inclusion or omission of a particle (e.g., the probability of occurrence/omission of the post-nominal particle was the same for all nouns and in all syntactic contexts). After exposure to the language, children and adults underwent a production test where their own usage of the variant particles was examined. The central finding was that, rather than reproducing the unpredictable variability in their own input, children showed *regularization*, so that their own productions were systematic: children either removed variation altogether (either by always using the particle or by dropping it altogether) or (more rarely) by conditioning particle use on noun type or sentence position (including the particle only after the object nouns, or only in transitive and not intransitive sentences). In contrast, adults reproduced the unpredictable variability present in their input and matched the frequency usage of the particle (a behaviour known as probability matching). Austin, Newport, and Wonnacott (2006) found similar results in an artificial language with two competing particles (rather than presence/absence variability), although here children almost always regularized by boosting the frequency of the more frequent form in the input.

[1]Hudson Kam and Newport (2005) refer to these words as “determiners”, although they were meaningless and served no discourse function, hence our preference for the more neutral term “particles”.

Other work using similar paradigms suggests that adults can also be induced to regularize, at least in some circumstances. Hudson Kam and Newport (2009) found that both adults and children showed regularization when variation was more complex (where there was one frequent particle and multiple low frequency alternatives), although adults still regularized less than children. Wonnacott and Newport (2005) looked at the learning of adults exposed to artificial languages containing unpredictable variation in either particle usage or word order and found more regularization with both types of variation in sentences involving novel (i.e., untrained) lexical items. Smith and Wonnacott (2010) demonstrated that even when adults appear to be probability matching, they may actually show a small but measurable tendency for regularization. They looked at the behaviour of adults who were exposed to a semi-artificial language, with nouns borrowed from English, where novel post-nominal particles marking plurality alternated unpredictably between two competing forms. Virtually all participants produced variable output given this variable input, and the population's mean production frequency of the two particles matched that of their input, indicators which are normally taken to be diagnostic of probability matching. However, in a second phase of this experiment, the miniature languages produced by this initial group of participants were given to a second group of participants to learn as a new target language, and then the output of those participants was passed in turn to a third group of participants, and so on in an iterated learning design. This transmission process revealed a regularization bias in adult learners, sufficiently subtle to be hard to identify in individuals but revealed by the amplifying effect of filtering the language through multiple adult learners (see also Reali & Griffith, 2009; Smith et al., 2017). Interestingly, the regularization Smith and Wonnacott (2010) observed in adults often involved introducing lexical conditioning: participants maintained usage of the two variants, but differentially associated them with different nouns, a type of conditioning that has not been reported in experiments with children.

Some types of linguistic variability are also more prone to regularisation than others. Culbertson, Smolensky, and Legendre (2012) showed that adult learners given input exhibiting variable word order will over-produce harmonic orders in their own output, favouring orders where modifiers appear consistently before or after the head of a phrase, with some nonharmonic orders being particularly dis-preferred; children showed a similar pattern of effects, with a stronger overall tendency to regularize and produce harmonic orders (Culbertson & Newport, 2015).

Finally, just as adults may sometimes regularize to some extent, children do not always fully regularize. Wonnacott (2011) demonstrated that 6-year-olds learning a semi-artificial language similar to that in Smith and Wonnacott (2010) (i.e. two particles, one more frequent than the other[2]) did not generally show the strong, absolute regularization found by Hudson Kam and Newport (2009), although they did show weaker but measurable regularization. Wonnacott (2011) speculated that the lack of strong regularization in children could be due to greater ease of lexical access due to the use of semi-artificial materials (see also Hudson Kam & Newport, 2009; and Hudson Kam & Chang, 2009).

In sum, the artificial language literature provides evidence that when learners are presented with a language containing unpredictable variation, they do not veridically reproduce that unpredictability; instead, they impose some systematicity in their own productions. This tendency to regularize appears stronger in children than adults, at least where the two have been directly compared. The results suggest that *both* child language acquisition and learning and usage of languages by adults might play a role in removing *unpredictable* variation from natural language. This could explain the fact that variation in

[2]We refer here to the Generalist Language in Wonnacott (2011) Experiment 1.

natural language is generally conditioned. But does this literature have implications for understanding the acquisition of naturally occurring sociolinguistic variation? It is worth noting that the seminal experiments by Hudson Kam and Newport (2005) were not designed to explore the learning of stable sociolinguistic variation; rather, their unpredictable input languages were designed to mimic the type of language produced by late language learners, i.e., individuals who are incorrectly using linguistic structures inconsistently in obligatory contexts (Newport, 1990; Sorace 2000). This is a type of variability that learners should not actually reproduce if they are to end up with a correct, native-like grammar. The finding that children did not acquire and reproduce these inconsistencies in the artificial languages was in line with the findings of Newport and colleagues (Newport, 1999; Ross & Newport, 1996; Singleton & Newport, 2004) who studied the language development of a profoundly deaf child ("Simon") who received American Sign Language (ASL) input only from his late-learner parents. In particular, Simon's parents were found to use a particular class of ASL morphemes inconsistently (around 70% of the time in contexts where it would be obligatory in native ASL), whereas Simon largely regularized away this variation and used the forms very consistently (around 90% of the time in obligatory contexts), despite receiving no other linguistic input (see Ross & Newport, 1996, for similar findings with deaf children learning ASL from hearing parents).

Turning to consider the learning of sociolinguistic variation, there is an obvious question: does the bias to regularize, which has been framed in terms of an advantage in "improving" inconsistent input, actually work *against* the learning of probabilistic, socially conditioned variation? Could regularization cause learners to eliminate the variation they encounter, rather than learn the conditioning contexts? There is some evidence that sociolinguistic variability can make acquisition slower, such that, for example, Chilean Spanish-speaking children take longer to acquire the variable plural marking of their

language (plural /-s/ is sometimes produced and sometimes omitted) relative to Mexican Spanish-speaking children learning the consistent plural marking of their language (plural /-s/ is categorically produced) (Miller & Schmitt, 2012). None of this work has, however, considered the presence of strong regularization in early child language, except for Habib's (in press) study of the Arabic rural /q/ [ʔ] variants discussed above: in this study, children at the youngest ages (prior to 8) used one variant categorically (the /q/ variant, more associated with female usage in adults and older children), moving to the adult-like, gendered usage only later. However, it is again difficult to know the extent to which this pattern reflected the input, for example, if early input came primarily from female caregivers and little usage of [ʔ] was encountered. Thus, it is not clear that regularization does occur when children learn "natural" linguistic variation.

Artificial language learning can also be used to look at the learning of conditioned variation. Hudson Kam and Newport (2009) showed that adults could successfully learn a complex artificial language where particle usage was conditioned on the noun, even though they showed regularization in a matched artificial language where particles and nouns were equally frequent but used inconsistently. Austin et al. (2006) and Wonnacott (2011) also showed that children could learn lexically conditioned variation in an artificial language.

Note that, in those experiments, the conditioning of particles was deterministic. Only one artificial language experiment to date has looked at the learning of probabilistic conditioned variation (Hudson Kam, 2015). Hudson Kam (2015) reports two experiments, one with adults and one with children (using somewhat different languages and methods), each comparing the learning of an artificial language in which the particles were used inconsistently (i.e., exhibiting unpredictable variation) with the learning of a matched language where usage of the particles was probabilistically conditioned on syntactic position. Results from adults confirmed that they reproduced unconditioned variation, as in the original

experiments (although, note that only absolute or near-absolute regularization was considered in these experiments, rather than more subtle reductions in variability as tested for in Smith & Wonnacott, 2010), and also showed that they were also able to approximate the conditioned patterns of usage in their input (they used particles more often in subject position than object position when this occurred in their input language, and vice versa). Children again showed strong regularization given inconsistent input, but importantly regularized the conditioned language less. Interestingly, this was despite the fact that children did not appear to have learned the actual conditioning in their input (they did not use particles more with subjects than with objects, despite that occurring in their input). Instead, there was some evidence that, at least some children, were imposing an alternative pattern of conditioning, whereby they were more likely to use particles in intransitive rather than transitive sentences. One possibility discussed by Hudson Kam (2015), is that children had picked up that the variation is conditioned but were initially tracking over the wrong variables, as has also been observed in the sociolinguistic literature (Labov, 1989).

The results of Hudson Kam (2015) suggest that, when languages contain probabilistic conditioning cues, variation may be more learnable. However, it has not yet been demonstrated experimentally that children can acquire the actual patterns of conditioning which are present in their input, nor how this learning interacts with regularization when it occurs. In addition, while Hudson Kam (2015) focused on a language-internal cue (syntactic conditioning), no experiment to date has looked at a language-external social cue. This is important since, as discussed in section 1.2, it is particularly difficult to establish through observations of naturalistic linguistic behaviour whether children can learn such cues. In the current study, we use artificial language learning techniques to explore whether learners can acquire variation conditioned on the sociolinguistic cue of speaker identity. The ability to track association between different speakers and their usage of different forms is necessary in

order to acquire any conditioning which is indexed by speaker identity, such as gender, social status, or ethnicity. In these experiments, we consider whether both adult and child learners can acquire this type of conditioning in an artificial language, and also explore the interaction with their bias for regularization.

1.4. The present study

Following Wonnacott (2011), we used *semi-artificial* languages, adapted for use with young children. This version of artificial language learning minimizes vocabulary learning during the experiment, thus, any age differences in the learning of morphosyntax cannot be due to differences in word learning ability (where adults outperform children). Furthermore, as shown by Wonnacott (2011; see also Wonnacott, Samara, Saul, & Smith, in preparation), semi-artificial languages reduce children's otherwise strong tendency to eliminate variation entirely, allowing us a better chance of seeing whether children can reproduce conditioned variability from their input.

In three experiments, 5- to 6-year-olds and adults were taught short sentences which were each produced by one of two speakers (one male and one female speaker). Each sentence consisted of a verb, one of 16 nouns (borrowed from English: e.g., *fox*, *cat*, *cow*), and one of two meaningless particles (e.g., *tid*, *kem*) which obligatorily followed the nouns. In Experiment 1, we sought to establish whether children and adults could learn without explicit instruction (i.e., incidentally) a pattern of sociolinguistic variation where the conditioning of particles around speaker identity was deterministic (e.g., the male speaker always used *tid* and the female speaker always used *kem*). Our tests probed production (where participants are asked to produce particles on behalf of a particular speaker) and comprehension (where they indicate which utterances are more appropriate for the different speakers). In Experiment 2, we repeated the experiment such that conditioning on speaker

identity was (more naturalistically) probabilistic (e.g., the male speaker used *tid* 75% of the time and *kem* 25% of the time; the female speaker used 75% *kem* and 25% *tid*). In both experiments, in addition to exploring participants' knowledge of correct/more-speaker-appropriate particle usage, when analysing production data we tested for the presence of any regularization, either through eliminating one particle, or through conditioning particle use lexically. Since these latter analyses suggested interesting differences between children and adults, in Experiment 3, we sought to explore whether these same patterns occurred even when there was no conditioning around speaker identity: we did this by training participants on languages where each of the two speakers used both particles equally.

We also used (i) tests of verbal and nonverbal short term and working memory ability to ensure that our participant groups across experiments were balanced in working memory capacity, and (ii) a post-experiment "structured interview" (see Appendix A for list of questions) to obtain subjective reports of participants' awareness of the relationship between particle usage and speaker identity. The relationship between explicit awareness and artificial language learning performance is not well understood, although it has received some interest in recent years (e.g., Hama & Leow, 2010; Leung & Williams, 2011; Williams, 2004, 2005). For sociolinguistic variables, the question of learner awareness is potentially interesting since, at least by adulthood, speakers are generally aware of the social values associated with different variants.

2. Experiment 1

2.1. Method

2.1.1. Participants

Participant details are show in Table 1. Thirty 5-6 year-old children and 30 adults took part in the study. Half (in each age group) were monolingual native English speakers

and the other half spoke another (or more than one) language in addition to English[3]. Participants had no known hearing, language, or speech disorders and were matched for standardized memory performance across experiments (adults: all *p*s > .510; children: all *p*s > .109)[4]. All-but-two participants completed the experimental sessions on four consecutive days; one child and one adult completed them over 5 days. Children were recruited from Year 1 and 2 of public primary schools in Coventry, Warwickshire, Greater London, Hertfordshire, and Scotland[5]. Children were rewarded with stickers and a certificate for their participation.

---

[3]Artificial language learning experiments often use only monolingual English speakers as participants, but this restriction is difficult for recruitment in Greater London. In addition, it is potentially of interest to see if bilinguals differ in their sensitivity to speaker identity cues, due to their greater experience in associating different individuals with different linguistic forms. We therefore aimed to recruit equal numbers of bilinguals and monolinguals for each condition and language group was initially included in all of the analyses to investigate any systematic differences between monolingual and bilingual participants. We found that there was no significant effect or interaction with language group, therefore, this factor was dropped from the analyses reported here.

[4]It has been previously suggested (Hudson Kam & Newport, 2009; see also Hudson Kam & Chang, 2009) that memory limitations could play a role in inducing regularization, potentially explaining why children regularize inconsistent input more than adults. To test this view, Perfors (2012) used a battery of measures to capture individual differences in memory ability among adults and assessed whether these predicted their differential performance on an artificial language learning task. In a similar vein, we also carried out a series of exploratory Pearson's correlations between children and adults' memory performance and the entropy scores reported in sections 2.2.5.1, 3.2.4.1, and 4.1.4.1. Although one must always be cautious about inferring from a null result, similar to Perfors (2012), we consistently found no significant relationship between the extent of regularization and any of the memory tests we administered in any of the three experiments.

[5]Due to differences in school-age entry between England and Scotland, child participants living in Scotland were recruited from Year 2 (referred to as Primary 2 in Scotland). All child participants recruited from schools in England were attending Year 1.

Written consent was obtained from their parents prior to the first session, and verbal consent was obtained from the children at the beginning of each session.

Adult participants were recruited through a university participant panel. They provided written consent at the beginning of the first session and were paid for their participation.

2.1.2. Input language and stimuli

Participants were exposed to a set of input sentences (n = 8), each repeated 32 times across the four experimental sessions. Every sentence described a picture of two cartoon animals and was three words long. The form of the sentence was *glim NOUN PARTICLE,* where *glim* was a novel verb meaning "THERE-ARE-TWO", NOUN was the relevant English label for the animal (e.g., "*pig*"), and PARTICLE was one of two particles (particle 1 or particle 2) and carried no semantics. Particles were monosyllabic nonwords (*bup*, *dak*, *fod*, *gos*, *jeb*, *kem*, *pag*, *tid*, *wib*, *yav*); for each participant we selected 2 at random from a set of 10, to minimize the impact of preferences for certain particles or certain particle-noun associations. Nouns were drawn from a set of 16 English nouns (*bear*, *cat*, *cow*, *dog*, *elephant*, *fox*, *giraffe*, *horse*, *lion*, *monkey*, *mouse*, *pig*, *rabbit*, *sheep*, *squirrel*, *zebra*; mean age of acquisition 4.24 years, *SD* = 0.61, Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) with 8 nouns used during the training and 8 reserved to be used as novel nouns during testing (random by-participant assignment of nouns to the trained/withheld sets).

Auditory stimuli (sentences in the semi-artificial language) were recorded by two speakers, speaker 1 (a female RP-English speaker) and speaker 2 (a male RP-English speaker).

Audio stimuli were recorded as separate words with appropriate prosody for their position in the sentence using a Sony PCM-M10 digital audio recorder. Peak amplitude was

normalised using Audacity (http://audacity.sourceforge.net/). Clipart pictures of the 16 nouns were selected from free online clipart databases, and were edited to generate pictures that showed pairs of items (e.g., two lions). Two cartoon pictures of a male and female face (with speech bubbles) were created to enhance speaker identity cues. The experiment ran in the Exbuilder software, a custom built experiment program written at the University of Rochester. This software carried out the by-participant vocabulary assignment and concatenated sound files to create the sentence stimuli in training and testing.

2.1.3. Procedure

Participants were tested individually in a quiet space in their school (children) or in a laboratory (adults), using identical instructions. Tasks were administered in a fixed order. Each session lasted approximately 20 minutes.

2.1.3.1. Semi-artificial language learning task

Participants were told that they were going to meet Henry and Katie who speak another language and that their task was to try to learn to speak like them. They completed four experimental sessions held on four consecutive days (notwithstanding two exceptions listed above). The tasks completed in each session, and the order in which they were administered, are summarised in Figure 1 and described in more detail below. Briefly: Language training occurred on all four sessions. A production test probing learning of the association between speaker and particle usage was administered on day 1 and day 4. A two-alternative forced choice task assessing sensitivity to the speaker identity cues was administered on day 4.

2.1.3.1.1. Noun practice

Participants first practiced saying the names of all 16 pictures. On day 1, they viewed a picture (e.g., one lion) while hearing its English name ("*lion*") and were required to say the name aloud. Names were spoken by the (female, non-RP) experimenter to avoid confusion with the critical male/female voice manipulation that predicted particle usage later in the experiment. Subsequently, participants viewed the same pictures and were asked to produce their names without prompts. This second naming task was repeated at the beginning of all subsequent sessions (days 2, 3, and 4) to ensure that participants labelled the pictures correctly (some pictures were consistent with more than one possible label; e.g., *rabbit*, *bunny*). When incorrect labels were provided, participants were corrected.

2.1.3.1.2. Sentence training

Participants saw a picture of the noun (e.g., two lions) in the centre of the screen and a cartoon picture of the male/female speaker with a speech bubble in the bottom left corner of the screen (Figure 2). They heard a sentence describing the picture (*e.g., glim lion bup*) and were asked to repeat it aloud. If any element of the sentence was mispronounced, feedback was provided ("*Almost, this one was "glim lion bup". Can you say that?"*). There were 8 repetitions per session (4 in male, 4 in female voice) of each of 8 nouns, for a total of 64 trials in each sentence training block. Critically, the occurrence of the two particles during sentence training was fully predictable according to speaker identity: speaker 1 used particle 1 100% of the time and speaker 2 used particle 2 100% of the time. There were no other cues to particle usage (e.g., facing direction of the animal). In sum, each particle was equally frequent and equally associated with each noun both in each session and across the language, but only produced by one of the two speakers.

2.1.3.1.3. Production test

On each trial, participants saw a picture (e.g., two lions), heard the first word of the sentence ("*glim"*) spoken by the male or female voice (accompanied by the cartoon picture of the speaker), and were asked to complete the rest of the sentence on behalf of Henry/Katie. If the noun was produced incorrectly participants were given corrective feedback (e.g., "*Good try, but this one is a lion, not a tiger*") and asked to say the sentence again using the correct noun. These trials were scored as incorrect unless the nouns produced were synonymous with the target noun (e.g., *bunny* was accepted in place of *rabbit*). No corrective feedback was provided regarding the sentence-final particles. If no particle was produced (e.g., "*glim lion*") after approximately 10 seconds, the experimenter enquired whether the participant was ready to move on to the next trial. There were 64 trials presented in a semi-randomized order precluding consecutive repetitions of the same noun. Thirty-two production trials tested performance on 4 of the 8 trained nouns (each noun encountered 8 times, 4 times with the female speaker prompt and 4 times with the male speaker prompt). The remaining 32 tested performance on 4 novel nouns (each presented 4 times with the female speaker prompt and 4 times with the male speaker prompt). Different novel nouns were used on days 1 and 4.

2.1.3.1.4. 2 Alternative Forced Choice (2AFC) test

On each trial, a picture of the noun (e.g., two lions) appeared centrally at the top of the screen. One second later, the cartoon picture of either the male or the female face (with a speech bubble) appeared in the bottom left corner of the screen and produced the appropriate (i.e., involving the particle used by that speaker) or inappropriate (i.e., using the other particle) sentence to describe the picture. The same cartoon face (with a speech bubble) then appeared in the bottom right corner of the screen and produced the other possible (i.e., appropriate/inappropriate) sentence. Participants were asked to click on the picture of the speaker who produced the sentence that described the picture "best". The order in which the

appropriate and inappropriate descriptions occurred was randomised across trials. Participants were tested on all 8 of the trained nouns and all 8 novel items that had been used in the production test. Half of the items in each set (i.e., trained vs. novel) were heard in the female voice and the other half were heard in the male voice.

2.1.3.2. Other tests

The Word Recall subtest of the Automated Working Memory Assessment *(*Alloway, Gathercole, Kirkwood, & Elliott, 2008), was administered in session 2 to assess verbal short-term memory ability. Mazes Memory and Backwards Digit Recall task from the same standardized memory test were administered in session 3 to assess nonverbal short-term memory and working memory ability, respectively.

2.1.3.3. Awareness questionnaire

At the end of the fourth session, all participants completed a verbal questionnaire probing explicit awareness of the association between speaker and particle usage (Appendix A). The same questions were administered to children and adults, although the former were prompted further, if necessary. Answers to this questionnaire were coded such that participants who did not indicate awareness of any patterns within the language received a score of 0, and those who explicitly mentioned that the 2 different voices used the particles in different ways received a score of 1.

2.2. Results

2.2.1. Statistical approach

Raw data from this project and analysis scripts are available on the Open Science Framework, http://osf.io/g67pe. The final R pub script can be found here:

http://rpubs.com/AnnaSamara/248957. For each of the analyses reported in sections 2.2.2 - 2.2.3 and sections 3.2.1 - 3.2.2, the dependent variable was binary (i.e., whether the particle in the response was correct/incorrect) and this data was analysed using logistic mixed effects models (Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Quené & van den Bergh, 2008) in the package *lme4* (Bates, Maechler, Bolker, & Walker, 2015) for the R computing environment (R Core Team, 2012). These models allow binary data to be analysed with logistic models rather than proportions, as recommended by Jaeger (2008). To avoid anti-conservative conclusions (Barr, Levy, Scheepers, & Tily, 2013), we specified a full random effect structure in our models, including intercepts for subjects and by-subject random slopes for all within-subject factors and their interactions. These models converged with Bound Optimization by Quadratic Approximation (BOBYQA optimization; Powell, 2009). In all models, predicting variables (including discrete factor codings) were coded as centered numerical predictors to reduce the effect of collinearity between main effects and interactions and in order that main effects were evaluated as the average effects over levels of the other predictors.

For each of the analyses reported in sections 2.2.5, 3.2.4, and 4.1.4, the dependent variables of entropy and mutual information were continuous and this data was analysed using *t* tests or mixed ANOVAs. However, since the data were severely skewed and did not meet the criteria for parametric inferential statistics, we computed lower and upper bootstrapped 95% confidence intervals of the mean(s) for all *t* tests and boostrapped the *p*

values using the R functions bwtrimbt (2-way ANOVAs) and bwwtrimbt (3-way ANOVAs)[6] (Wilcox, 2012). Categorical data were analyzed using chi-square or Fisher's exact tests.

2.2.2. Production data

Our main analyses excluded trials where children did not produce a correct sentence of the form *(glim) NOUN PARTICLE* where the particle was clearly identifiable as one of the two occurring in the child's input. Errors included producing an incorrect noun, producing English or random utterances, or producing something in place of the particle which could not be identified as one of the two particles occurring in the child's input (e.g. ''*glim mouse glim*", ''*glim squirrel sleep*") or producing a blend of the two particles "*glim cat jem*", where the two input particles were "jeb" and "kem"). Mispronunciations which were identifiable as one of the two particles (e.g. a single phoneme substitutions as in jeb -> jep) were not excluded. All nonparticle responses which entailed a single-phoneme substitution were evaluated by two independent coders (the first author and a trained research assistant) and, when possible, were recoded to match one of the two particles (e.g. jep -> jeb). Inter-rater agreement on these items was very high (ICC = .95).

The proportion of valid sentences produced provides a baseline measure of learning. This is therefore also reported and analyses were included in order to investigate how these errors were distributed across testing days and noun types.

2.2.2.1. Baseline

[6]Note that bwtrimbt and bwwtrimbt do not generate bootstrapped *F* statics, thus, only bootstrapped p values are reported.

Overall, 9.77% of children's data was excluded according to the criteria above. Two of the thirty children produced no usable data on day 1 and did not contribute to the analyses on choice of particle (section 2.2.2.2). A logistic mixed effects model predicting the occurrence of an excluded sentence, with day (day 1, day 4) and noun type (old, novel) as fixed effects found no effect of either day (b = -5.69, SE = 6.73, z = -0.85, p = .398) or noun type (b = -0.26, SE = 1.77, z = -0.15, p =. 884) and no interaction between day and noun type (b = -2.23, SE = 3.83, z = -0.58, p = .560). Thus, the proportion of excluded responses was stable across days and noun types.

2.2.2.2. Choice of particle

Participants' use of correct (i.e., speaker appropriate) particles is shown in Figure 3. On inspection, the proportion of correct responses appears to be above chance (50%), and adults generally outperform children. Both of these observations were statistically confirmed.

The effect of age on performance was tested in a logistic linear mixed effect model predicting accuracy with age group (children, adults), day (day 1, day 4), and noun type (old, novel) as fixed effects. The model showed a significant intercept (b = 4.83, SE = 0.39, z = 12.32, p < .001) which suggests that mean performance was significantly greater than chance (50% accuracy)[7] and a significant effect of age group (b = -3.63, SE = 0.71, z = -5.13, p < .001) (adults: M = .97, SD = 0.16; children: M = .84, SD = 0.37). No other effects were significant, all ps > .312.

[7] In mixed effect model analyses, conclusions re: chance performance (typically established via one-sample t tests) can be established by examining the significance of the model's intercept compared to the value expected under the null hypothesis, in this case equal use of the two particles for each speaker which would yield 50% correct trials, or log-odds of a correct response of 0.

Both children and adults showed strong learning of the deterministic speaker identity cues and used this information in the production task, although the effect was greater for adults than children. For both age groups, speaker appropriate particle usage was already present after one session and held equally for trained nouns and novel nouns.

2.2.3. 2AFC data

Proportions of correct responses in the 2AFC task (i.e., choosing the sentence with the speaker appropriate particle) as a function of noun type are shown separately for children and adults in Figure 4. On inspection, it is clear that 2AFC performance was better than chance across age groups and noun types, and better in adults than children.

This was confirmed in a logistic mixed effect model with age group (adults, children) and noun type (old, novel) as fixed effects. There was a significant intercept (b = 6.89, SE = 1.40, z = 4.91, $p < .001$) showing that mean performance was significantly greater than chance, and a significant effect of age group (b = -4.55, SE = 1.72, z = -2.65, p = .008) (adults: M = .99, SD = 0.11; children: M = .81, SD = 0.39). There were no other significant effects, all ps > .485.

In sum, as in the production data, both children and adults made speaker identity appropriate particle choices more frequently than expected by chance, with adults producing more correct responses than children. This held equally across trained and untrained nouns.

2.2.4. Questionnaire data

Responses on the post-experiment interview were binary coded to indicate whether participants were explicitly aware of the association between particle usage and speaker identity. To be classified as “aware”, participants had to verbally report (in response to either question 3, 4, or 5; see Appendix A) that they had noticed the association between particle

usage and speaker identity. As expected, children's and adults' "aware" responses differed in their level of sophistication, however, they all made mention of the face/head/gender of the speaker (e.g., "*I chose based on the gender: Katie said bup and Henry said gos*"). Examples of "unaware" responses often included that particle choice was "*fair*", due to "*careful listening of what Henry/Katie said*" or was based on an irrelevant aspect of the pictures (e.g., the direction that animals were facing, which was actually random). All responses were coded by the first author and an independent coder, producing an ICC of .97 for children and 1.00 for adults.

All adults and 40% of the children reported that particle usage was conditioned on the identity of the speaker (see Table 2). To investigate whether significant learning performance was always accompanied by explicit awareness, for children, we repeated the analyses reported above excluding participants who were classified aware. These analyses, reported in Appendix B, suggested that both production and 2AFC performance remained significant when "aware" children were excluded.

2.2.5. Regularization analyses

2.2.5.1. Overall regularization

One method by which participants might reduce the unpredictability of particle usage in their own productions is to boost the frequency of one of the two particles at the expense of the other. To capture this type of regularization, we computed the total entropy of each participant's particle use, given by:

$$Entropy = -\sum P(i)\ log_2 P(i)$$

where the sum is over the two possible particle choices, and *P(i)* is the frequency of particle *i* in a participant's productions.

Entropy measures how variable a participant's productions are: entropy of 0 corresponds to a participant who consistently uses a single particle, and entropy is at a maximum (entropy = 1) when both particles are used equiprobably (as in the input). It is also possible that participants may regularize differently across familiar (i.e., trained) and novel items (Wonnacott & Newport, 2005): we therefore calculated total entropy separately for each participant in each session for both old nouns and novel nouns, and across all productions combined (i.e., old + novel nouns)[8]. Figure 5 (first row, left panel) shows the mean entropy for each of these measures for adults and children in each session; the dotted line shows the average entropy predicted if participants produced particles at random, with each particle being produced with equal probability on any given trial (0.988, obtained via simulation).

Beginning with whole-language entropy, we used a mixed ANOVA with age group (adults, children) as a between-subjects factor and day (day 1, day 4) as a within-subject factor. This showed a significant difference in mean whole-language entropy (mean = 0.83; 95% CIs = [0.75, 0.91]) relative to the previously computed chance entropy (0.988) and a significant main effect of age group (p = .014), indicating that children's productions have lower entropy (mean entropy = 0.71) (i.e., they are more regular) than those of adults (mean entropy = 0.97). There was also a main effect of day (p = .041), reflecting greater

[8] Note that whole language entropy is not equivalent to the average of entropy for old and novel nouns separately, therefore, separate analyses were carried out. For example, if a participant used one particle consistently for old nouns and the other consistently for new nouns, both old and new nouns considered separately would have entropy of 0, but whole-language entropy would be 1.

regularization on day 1 (mean entropy = 0.80) than day 4 (mean entropy = 0.88), and no interaction between day and age group (p = .101).

We were also interested in establishing whether each age group showed above chance regularization on each day, thus, we also compared the whole-language entropy of each group on each day to chance. This revealed that children's entropy was below chance on both days (day 1: 95% CIs = [0.49, 0.80]; day 4: 95% CIs = [0.60, 0.89]); adults' entropy was not significantly below chance on day 1 (95% CIs = [0.85, 1.00]) and was actually significantly *greater* than chance on day 4 (95% CIs = [0.999, 1.00]). This last result reflects the fact that, by the final day, adults strongly condition particle usage on speaker, and therefore are using the two particles much more evenly across the language than would be predicted by chance usage.

To see if there was any difference in regularization across old and novel nouns, we also ran a second analysis using the data from old and novel nouns and including noun type (old, novel) as a within-subjects factor along with age group (children, adults), day (day 1, day 4) and their various interactions. This model confirmed the main effect of age group and day and the absence of an interaction between these factors, but also found a main effect of noun type (p = .038), suggesting that children and adults showed significantly more regularization for novel nouns (mean entropy = 0.82) than old nouns (mean entropy = 0.86); there were no other significant effects, $p > .065$.

As a subsidiary analysis, we also evaluated whether each individual participants' entropy was reduced relative to the input language *to a greater extent* than would be predicted by chance (i.e., with each particle being produced with equal probability on any given trial, with the total number of trials for each participant being matched to the number of valid data points they actually produced; this was obtained via simulating 1000 "random" production datasets for each participant, and observing the proportion of those datasets that

had entropy less than or equal to the actual entropy of that participant's data). We focused on the whole language measure of entropy and took $p < .05$ as our level of significance, and simply counted how many participants were producing significantly more regular output than expected by chance. Classified in this way, 13/28 children (recall that two did not produce either particle) and 2/30 adults regularized on day 1 (a significant age difference: $\chi^2 = 11.94$, $p = .001$); 11/30 children and 0/30 adults regularized on day 4 (a significant age difference: $\chi^2 = 13.47$, $p < .001$)[9]. This suggests that there were no significant qualitative changes in

[9] Another question of interest is whether individual participants produced the same patterns of regularization on days 1 and 4. That is, for a given participant, how different were his/her rates of particle production across days? We addressed this question by computing a measure of participants' self- consistency across days and comparing it against the level of self-consistency expected under the null hypothesis that participants are *not* self-consistent across days (i.e., participants' particle usage on day 4 no more resemble their day1 particle usage than expected by chance). For every participant, we calculate a difference score based on their particle usage on day 1 and day 4; this difference score can simply be calculated on the overall usage of particle 1 on both days, or calculated on noun-specific proportions. For example, a participant who uses particle 1 at approximately the same frequency on day 1 and day 4 (either averaging across nouns, or per noun) will have a low difference score; a participant who changes more dramatically from day 1 to day 4 (e.g., going from never to always using particle 1) will have a high difference score. We then compare the mean of the veridical difference scores to a distribution of mean difference scores generated by shuffling day 4 data across participants (e.g. re-attributing the data produced by participant 1 on day 4 to participant 2 and so on; we generate 1000 such random re-assignments to obtain a distribution of mean difference scores), which captures the null hypothesis that participants are no more self-consistent across days than we would expect by chance, given the observed distribution of responses on day 1 and day 4. We then obtain a p value for the observed level of self-consistency by calculating the proportion of samples from this null distribution that exhibit mean difference scores which are equal to or lower than the veridical mean difference score; if less than 5% of the random samples have equal or lower mean difference, we can reject the null hypothesis at $p < .05$ and conclude that participants are more self-consistent than we expect by chance. This analysis produced

children's regularization status due to the additional training received between day 1 and day 4. It is worth noting that some of these significant regularizers were *absolute* regularizers (i.e., only used one particle, producing output with entropy = 0). 6/13 children on day 1, 1/2 adults on day 1, and 5/11 children on day 4 were absolute regularizers, in other words the majority of regularizers were *not* absolute regularizers, even for children; rather, they produced output which featured both particles, but featured one particle significantly more often than would be expected if they were probability matching.

2.2.5.2. Lexical conditioning

Previous work (e.g., Hudson Kam & Newport, 2009; Smith & Wonnacott, 2010) suggests that, in addition to regularizing across the language, participants might also regularize around lexical items, i.e., condition their particle use on the noun preceding the particle, producing particular particles more frequently with particular nouns (despite particles being used equally frequently with each noun in the input). For example, in Hudson Kam and Newport's (2009) study, a few adult participants who imposed regularities on the inconsistent target language did so by systematically associating individual determiners with

---

clear evidence that children were more self-consistent than expected by chance in all 3 experiments, regardless of whether we look at overall or noun-specific particle usage (all *p*s <.002). For adults, there is only clear evidence of levels of self-consistency greater than chance in Experiment 3 (*p*s < .011); for experiments 1-2 these tests indicate no more self-consistency than expected by chance (experiment 1: *p*s > .999; experiment 2: *p*s > .094). However, adults performed very homogeneously in the presence of speaker-based conditioning (i.e. in Experiments 1 and 2), which means that it is difficult to demonstrate self-consistency greater than chance under our conservative null model: for example, in Experiment 1, most participants produce particle 1 on 50% of all trials on day 4; consequently, shuffling across participants yields a difference scores essentially identical to the observed difference score, which means that we cannot reject the null hypothesis in these cases.

individual lexical nouns. We sought to systematically explore this type of regularization by considering entropy for each participant's productions per lexical item. One possible metric would simply be to consider the conditional entropy of particle use given the noun being marked, which is simply the average entropy of particle use for each noun considered separately (where the average is weighted by noun frequency). However, this measure fails to differentiate between regularization at the lexical level (which yields low conditional entropy) and regularization across the language as a whole: participants who predominantly used one particle in most of their productions would have low entropy for the whole language and, as a consequence, for each lexical item. To isolate purely lexically based regularization, we therefore computed the *Mutual Information (MI)* of particle use and lexical identity, consisting of:

Mutual Information (MI) = total entropy – conditional entropy of particle given noun

Lexical MI of 0 indicates that a participant shows no lexically based regularization (i.e., the variability of each noun mirrors that of the language as a whole), while scores of 1 would indicate perfect lexical conditioning of particle use on the noun (i.e., the language as a whole would exhibit high variability, but each noun would exhibit zero variation in particle use). Figure 5 (second row, left panel) shows the mean MI for each of these measures for adults and children in each session; the dotted line shows the average predicted MI if participants produced particles at random, with each particle being produced with equal probability on any given trial (MI = 0.080, obtained via simulation). It can be seen that participants have MI scores very close to 0, indicating the absence of lexical conditioning. A mixed ANOVA with age group (adults, children) as a between-subjects factor and day (day 1, day 4) as a within-subject factor showed that the overall lexical MI was significantly less than

chance (mean MI = 0.02; 95% CIs = [0.01, 0.02]); there was a significant effect of age group (p = .006), due to lower MI in adults (Mean MI = 0.00**)** than children (Mean MI = 0.03), no effect of day (p = .240), and no interaction between age group and day (p = .362). In spite of the age difference in terms of MI, both groups' lexical MI was significantly less than chance (adults: 95% CIs = [0.00, 0.01]); children: 95% CIs = [0.02, 0.04]). These results are due to the fact that both age groups were spreading their production of particles more evenly across the words than would be predicted by chance usage, due to their learning of the conditioning on speaker.

A further mixed ANOVA using the data from old and novel nouns with age group (children, adults), day (day 1, day 4), and the additional within-subject factor noun type (old, novel) confirmed the above pattern of results and further showed that there was no effect of noun type (p = .218), and no interaction with this factor (noun type by age group: p = .371; noun type by day: p = .145; noun type by day by age group: p = .215).

We also computed via simulation for each participant whether their MI differed from that predicted by chance (given both the number of each of the two particles they produced and the number of valid productions they generated). This was not the case for any child or adult participant, again indicating the absence of spontaneous lexical conditioning of variation in this experiment.

2.2.5.3. Summary of regularization data: Experiment 1

Children show greater regularization in particle usage across the language (i.e., greater reduction of entropy) than adults. The extent of regularization decreased over the course of exposure to the language, from day 1 to day 4. There was also evidence of greater regularization with novel nouns than with old nouns. No lexically based conditioning was seen in this experiment.

2.3. Summary and discussion

Six-year-old children and adults were exposed to and tested on a semi-artificial language using identical procedures. The critical feature of the language was that nouns were followed by one of two particles whose use was fully predictable from the identity of the speaker: sentences produced by the female speaker were 100% associated with particle 1 and sentences produced by the male speaker were 100% associated with particle 2. To assess learning, participants were administered a production task at the end of the first and last training session (days 1 and 4) and a 2AFC task at the end of the last training session (day 4).

Our results indicate that participants in both age groups rapidly picked up on the speaker-based conditioning of variation in their input: both groups of participants demonstrated above-chance performance in both the production and 2AFC tests, even on day 1. Performance on the production task improved from day 1 to day 4, suggesting a role for input quantity. Age effects were apparent in both tasks, with adults outperforming children, and, in fact, showing near ceiling performance in both tasks. In addition, the questionnaire data revealed that all adults in Experiment 1 could explicitly verbalize the relationship between speakers and particle use, while this was true for only half of the children. Children's overall performance remained above chance (indicative of having implicitly acquired the conditioned pattern of variation) even when these "aware" participants were excluded. We return to consider the relationship between learning and explicit awareness in the General Discussion. For now, the results suggest that, at least by age 5-6, children can pick up the linguistic forms produced by different speakers incidentally, even when they are unaware of the constraints in question.

The second series of analyses aimed to assess whether either age group showed evidence of regularization, as seen in previous studies (Hudson Kam & Newport, 2005, 2009;

Smith & Wonnacott 2010). We used analyses of entropy and MI to look both for regularization across the whole language (i.e., boosting the frequency of one particle over the other) or lexical regularization (i.e., producing particular particles more frequently with particular lexical items). Only the first type of regularization was seen and only in children. This age difference is similar to that reported by Hudson Kam and Newport (2005, 2009), although it is important to note that, in our experiment, the fact that children and adults are able to learn and exploit the conditioning of variation based on speaker should lead to the lower levels of regularization. Similarly, we saw less regularization on day 4, which could reflect better learning of the input patterns following greater exposure. There was also more regularization for novel than old nouns (mirroring the results from Wonnacott & Newport's 2005 experiment with adults and extending that finding to another age group). We leave further consideration of this findings for the General Discussion.

Returning to our key result, Experiment 1 demonstrates that 5- to 6-year-old learners (and adults) can track sociolinguistic variation conditioned on speaker identity. This is important since it suggests that the previously reported tendency to regularize–which is still present in our results–does not prevent the learning of variation when that variation is conditioned. This ability to track conditioning on speaker identity must underpin the ability to acquire any sociolinguistic variation which is indexed by speaker identity or group membership.

In this experiment, speaker identity conditions variant use deterministically. In natural languages, the social and stylistic constraints governing sociolinguistic variation tend to be probabilistic rather than deterministic (Kroch, 1989; Labov, 1989). It is therefore important to establish whether children and adults can also learn sociolinguistically constrained variation from such probabilistic input. We address this question in Experiment 2.

3. Experiment 2

3.1. Method

3.1.1. Participants

Thirty 5-6 year-old children and 30 adults without known hearing, language, or speech disorders participated in the study (Table 1). All participants completed the experiment on four consecutive days. Children were recruited from Year 1 and 2 of public primary schools in Coventry, Warwickshire, Greater London, Hertfordshire, and Scotland. Adults were recruited through a university participant panel and received monetary compensation.

3.1.2. Stimuli, design and procedure

The stimuli and procedure were identical to Experiment 1. In Experiment 2, particle usage during training was partially consistent such that the female speaker used particle 1 75% of the time and particle 2 25% of the time, whereas the opposite held true for the male speaker. The partial consistency applied across the entire training set and also for each noun, such that, in each training session, participants heard each noun 4 times in total (3 times with particle 1 and once with particle 2) in one of the voices, and vice versa in the other voice.

3.2. Results

Data analysis procedures and baseline criteria for data point inclusion were identical to those employed in experiment 1. ICC for nonparticle responses in children's data was .98.

3.2.1. Production data

3.2.1.1. Baseline

Overall, 8.35% of data for children did not meet the baseline. The proportion of rejected responses was stable across days and noun types: a logistic mixed effects model predicting the occurrence of an excluded sentence with day (day 1, day 4) and noun type (old, novel) as fixed effects found no effect of day (b = -2.43, SE = 2.78, z = -0.87, p = .383) nor noun type (b = -0.94, SE = 1.95, z = -0.48, p = .629), and no day by noun type interaction (b = -5.55, SE = 3.89, z = -1.43, p = .153).

3.2.1.2. Choice of particle

As in Experiment 1, we investigated children's and adults' ability to produce speaker-appropriate particles across days and noun types. As seen in Figure 6, there was some indication of differences between children's and adults' performance on day 4, with adults being closer to the trained frequencies on day 4. A logistic mixed effect model with age group (adults, children), day (day 1, day 4), and noun type (old, novel) as fixed effects was fitted to the proportion of majority-particle (i.e., speaker appropriate) produced. The model showed a significant intercept (b = 0.28, SE = 0.08, z = 3.55, p < .001), no significant effect of age group (b = -0.22, SE = 0.16, z = -1.38, p = .166), no significant age group by day interaction (b = -0.14, SE = 0.08, z = -1.78, p = .075), and a significant effect of day (b = 0.08, SE = 0.04, z = 1.99, p = .046). The only other significant effect in the model was the three-way interaction between age group, day, and noun type (b = 0.15, SE = 0.07, z = 2.20, p = .028).

We also considered task performance separately for children and adults. The model predicting adults' production of the majority particle as a function of day and noun type showed a significant intercept (b = 0.40, SE = 0.12, z = 3.26, p = .001), and a significant effect of day (b = 0.15, SE = 0.06, z = 2.34, p = .019) demonstrating that the proportion of speaker appropriate particles increased from day 1 (M = .54, SD = 0.50) to day 4 (M = .62,

SD = 0.48). Mean performance was not significantly greater than chance (50%) on day 1 (*b* = 0.17, *SE* = 0.14, *z* = 1.22, *p* = .221), but was significantly above chance on day 4 (*b* = 0.62, *SE* = 0.17, *z* = 3.64, *p* < .001). There was no significant effect of noun type (*b* = -0.04, *SE* = 0.08, *z* = -0.50, *p* = .618), or interaction between day and noun type (*b* = -0.09, *SE* = 0.05, *z* = -1.66, *p* = .098). A similar model fitted to children's production of the majority particle showed no significant intercept (*b* = 0.16, *SE* = 0.10, *z* = 1.52, *p* = .128), suggesting that mean performance was at chance levels. There was no significant effect of day (*b* = 0.00, *SE* = 0.05, *z* = 0.09, *p* = .926), noun type, *b* = -0.03, *SE* = 0.07, *z* = -0.43, *p* = .665, or day by noun type interaction, (*b* = 0.06, *SE* = 0.05, *z* = 1.16, *p* = .248). Thus, adult performance differed from chance on day 4 but not day 1, while children were at chance across both days. Note that neither group show a significant difference between old and novel nouns nor an interaction between noun type and day[10].

In addition to considering whether participants are above chance in their usage of the particles (which shows learning), it is also of interest to see whether participants are producing particles with each speaker with approximately the frequency seen in the input (i.e. probability matching for each speaker). Note that it would be possible for learners to introduce another type of regularization here, whereby they boost the frequency of the majority particle for each speaker, thus, making their productions more like the input in Experiment 1 (i.e., conditioning particle use on speaker identity). This type of conditioning

[10] Breaking down the significant three-way interaction in the overall model revealed a significant two-way interaction between day and age group for novel nouns, not old nouns. This suggests that for novel nouns, adults improved more than children, but for old nouns they did not. However, given that adults do not show a significant day by noun type effect, and that children are generally at chance, we find this difficulty to interpret.

did not occur. In fact, both children and adults undershot the speaker-specific probabilities in their input, i.e., they produced output in which variation was *less* conditioned on speaker identity relative to their input. To test whether participants were significantly undershooting, we compared the proportions of majority particle produced by children and adults against the trained odds ratio in the input, i.e., 3 (75%). The observed odds ratio (1.13; 53%) in the model of children's performance was significantly lower than 3 (z = -9.02, p < .001), and so was the observed odds ratio in the model of adults' performance, both on day 1 (observed odds ratio = 1.74 (54%), z = -6.66, p < .001) and day 4 (observed odds ratio = 1.63 (62%), z = -2.76, p = .006).

In sum, adults' production performance showed significant sensitivity to the speaker identity manipulation on the final day of testing, i.e., usage of the majority particle with the associated speaker was significantly greater than chance (50%). They nevertheless significantly undershot the probabilities of the input (62% usage of the majority particle for each speaker, compared with 75% in the input). Adults' performance on day 1 and children's performance on both days did not differ from chance.

3.2.2. 2AFC data

Proportions of majority particle choice in the 2AFC task in Experiment 2 are shown in Figure 7. As in Experiment 1, we began by running a logistic mixed effect model to investigate differences in 2AFC performance as a function of age group (adults, children) and noun type (old, novel). The model suggested that mean performance was overall above chance (b = 0.40, SE = 0.09, z = 4.53, p < .001), and that adults' performance was not significant better (M = .62, SD = 0.49) than children's performance (M = .56, SD = 0.50) (b = -0.31, SE = 0.18, z = -1.76, p = .079). There were no other significant effects or interactions, ps > .280.

We also fitted separate model for adults and children to investigate performance in each age group. Both adults' and children' mean performance was significantly greater than chance (adults: $b = 0.58$, $SE = 0.14$, $z = 4.03$, $p < .001$; children: $b = 0.24$, $SE = 0.11$, $z = 2.19$, $p = .029$), and no main effect of noun type (adults: $b = 0.23$, $SE = 0.22$, $z = 1.08$, $p = .281$; children: $b = -0.07$, $SE = 0.19$, $z = -0.38$, $p = .701$). Thus, both children and adults picked speaker identity appropriate particles more frequently than expected by chance, an effect which did not differ significantly across trained and novel nouns.

3.2.3. Questionnaire data

Inter-rater reliability (ICC) for the assessment of explicit awareness in Experiment 2 was 100% for children and 93.33% for adults. Approximately half of the adult participants (12/30; 40%) but only a handful of child participants (3/30; 10%) were explicitly aware of the probabilistic association between particle usage and speaker identity. Adults' "aware" responses made mention of a partial association between speaker identity and particle usage (e.g. "*I associated wib more with Henry and gos more with Katie*", "*Generally she said yav more and he said bup more*"), although some were not 100% accurate (e.g., "*usually, when you switched from Henry to Katie, she produced bup instead of gos*"; "*Katie says jeb most of the time but for some animals she says wib more*"). "Unaware" responses made no mention of speaker identity. For example, some participants reported an incorrect frequency-based pattern ("*I noticed there were more "daks" than "kems*") while others reportedly noticed that particle alternation "*followed a sequence*".

As in Experiment 1, a series of analyses were carried out to investigate learning performance excluding those participants that were coded as "aware" (Appendix C). Table 3 also summarizes task performance in participants classified as "unaware". For child participants, statistical analysis demonstrated that exclusion of aware participants did not alter

any of the key findings reported earlier. For adult participants, on the other hand, sensitivity was not significant in either production or 2AFC performance (*p*s > .2) once "aware" adults were excluded.

3.2.4. Regularization analyses

3.2.4.1. Overall regularization

Again, we looked at entropy as a measure of how variable participants' productions were. Results are shown in Figure 5 (first row, middle panel). A mixed ANOVA on whole-language entropy (i.e., for old and novel nouns combined) with age group (adults, children) as a between-subjects factor and day (day 1, day 4) as a within-subject factor showed a significant main effect of age group ($p < .001$), reflecting lower entropy in children's productions than in adults, and a main effect of day ($p = .001$), reflecting greater regularization on day 1 than day 4. Unlike Experiment 1, in this experiment, there was also a significant interaction between age group and day ($p = .001$). Breaking this down showed that, for children, mean entropy on day 1 differed significantly from mean entropy on day 4 (mean difference = 0.34; 95% CIs = [0.18, 0.51]), whereas for adults, it did not (mean difference = 0.01; 95% CIs = [-0.004, 0.01]). Children's entropy was below chance (0.983, computed via simulation) both on day 1 (95% CIs = [0.19, 0.48]) and day 4 (95% CIs = [0.53, 0.80]), whereas adults' entropy across days was not significantly different from chance (95% CIs = [0.980, 0.99]).

To see if there was any difference in regularization across old and novel nouns, we also subjected the data from old and novel nouns to an Age Group by Day by Noun Type mixed ANOVA. This analysis confirmed the main effect of age group, day, and the interaction between them, but also found a main effect of noun type ($p = .002$) qualified by an interaction between noun type and age group ($p < .001$). There were no other significant

effects, *p*s > .127. Breaking down the interaction showed that children showed significantly more regularization for novel nouns (mean entropy = 0.45) than old nouns (mean entropy = 0.52) (mean difference = 0.06; 95% CIs = [-0.11, -0.01]) but adults did not show a significant difference for novel nouns (mean entropy = 0.95) and old nouns (mean entropy = 0.96) (mean difference = -0.01; 95% CIs = [-0.04, 0.02]).

As for Experiment 1, we also used simulation to calculate the number of adult and child participants who reduced entropy compared with the input language to a greater extent than would be predicted by chance. Classified in this way, 23/30 children and 3/30 adults regularized on day 1 (a significant age difference: $\chi^2 = 27.15$, $p < .001$); 18/30 children and 0/30 adults regularized on day 4 (a significant age difference: $\chi^2 = 25.71$, $p < .001$). Of these significant regularizers, the following were absolute regularizers (i.e., only used one particle): 12/23 children and 0/2 adults on day 1; 5/18 children and 0/2 adults on day 4. In other words, as in Experiment 1, many children and all adults who produced significantly regularized output nonetheless exhibited *some* variability; only for children on day 1 were the majority of regularizers producing absolutely regular data exhibiting zero variation.

Finally, it is also of interest to consider whether there was more regularization in this experiment than in Experiment 1. We chose to do this by comparing the number of significant regularizers (i.e., individuals who produced particles with below-chance entropy, as identified by simulation). Using this binary classification results in some loss of precision, however, the alternative of comparing entropy across experiments is less suitable in that it would potentially pick up on differences due to entropy being artificially higher than chance when participants are strongly conditioning on speaker. Instead, the point of interest is the extent to which entropy is below chance due to participants' tendency to regularize. The analyses showed that, for children, there were more regularizers in this experiment than in Experiment 1 on both days, though this was only significant on day 1 (13/28 in Experiment 1

vs. 23/30 in Experiment 2: $\chi^2 = 5.62$, $p = .018$); On day 4, the difference between regularizers in this experiment and Experiment 1 was not significant (11/30 in Experiment 1 vs. 18/30 in Experiment 2, $\chi^2 = 3.27$, $p = .071$). There was no significant difference for adults on either day 1 (2/30 in Experiment 1 vs. 3/30 in Experiment 2, $p = 1.000$, two-tailed Fisher-exact test) or day 4 (0/30 in Experiment 1 vs. 0/30 in Experiment 2).

3.2.4.2. Lexical conditioning

Lexical MI was computed as in Experiment 1; results are shown in Figure 5 (second row, middle panel) (recall that here high score indicates *greater* lexical conditioning). We first analysed whole-language MI (i.e. for old and novel nouns combined) in a mixed ANOVA with age group (children, adults) as a between-subjects factor and day (day 1, day 4) as a within-subjects factor. Mean whole-language MI was significantly different relative to chance MI (MI = 0.077, obtained via simulation) (mean MI = 0.13; 95% CIs = [0.10, 0.17]), however, there was also a significant main effect of age group ($p < .001$), indicating that adults' productions have higher MI (i.e., more lexically-based regularization) than those of children. There was also a main effect of day ($p = .017$), and a significant interaction between age group and day ($p = .003$). As shown in Figure 5, for adults, the decrease in MI from day 1 and day 4 was significant (mean difference = -0.15; 95% CIs = [-0.24, -0.07]), whereas for children it was not (mean difference = 0.01; 95% CIs = [-0.01, 0.03]). Breaking down the effect of day for adults showed that, although the effect was stronger on day 1, lexical MI was significantly greater than chance (0.077) on both days (day 1: 95% CIs = [0.21, 0.37]; day 4: 95% CIs = [0.09, 0.17]). In contrast, children's overall lexical MI was actually below chance (mean MI = 0.06; 95% CIs = [0.04, 0.076]), suggesting that they lexicalized less than would be expected by chance. This is possibly due to the fact that their overall regularization is high in this experiment.

To see if there was any difference in regularization across old and novel nouns, we also analyzed the data from old and novel nouns in an Age Group by Day by Noun Type mixed ANOVA. The analyses confirmed the main effect of age group, day, and the interaction, and found no main effect of noun type (p = .160) or any interactions with noun type (noun type by age group: p = .756; noun type by day: p = .115; noun type by age group by day: p = .267).

We also computed via simulation, for each participant, whether their MI score was significantly lower than that which would be expected by chance, given both the number of each of the two particles that they produced and the number of nouns that they produced. Classified this way, 0/30 children and 15/30 adults produced lexically conditioned output on day 1 (χ^2 = 20.00, p < .001); 1/30 children and 9/30 adults produced lexically conditioned output on day 4 (χ^2 = 7.68. p = .006).

As for entropy, we compared lexically based conditioning across experiments by comparing the number of significant regularizers. Recall that in Experiment 1 there was *no* lexically based conditioning. The number of regularizers in the two experiments differed for adults on both days (day 1: 15/30 in Experiment 1 vs. 0/30 in Experiment 2, χ^2 = 20.00, p < .001; day 4: 0/30 in Experiment 1 vs. 9/30 in Experiment 2, p = .002, two-tailed Fisher-exact test) but not for children on either day (day 1: 0/28 in Experiment 1 vs. 0/30 in Experiment 2; day 4: 1/30 in Experiment 1 vs. 0/30 in Experiment 2; p = 1.000, two-tailed Fisher-exact test).

3.2.4.1. Summary of regularization data: Experiment 2

Children showed greater regularization in particle use than adults. The extent of regularization decreased from day 1 to day 4, though this was only significant in children. As in Experiment 1, there was evidence of greater regularization with novel nouns than with old

nouns (though here the difference is significant only for children). For children only, comparison of the number of significant regularizers suggested more regularization in Experiment 2 than in Experiment 1, though this was only significant for day 1. This is presumably due to children's learning of the speaker-based conditioning in Experiment 1. Adults showed no evidence of significant regularization, however they did show a strong tendency to lexically condition particle use. This contrasts with Experiment 1 where adult participants successfully acquired speaker-based conditioning of variation and did not lexically condition. This tendency to lexically condition reduced with further training on the (lexically unconditioned) target language, although it was significant in both sessions. Children showed no evidence of lexically conditioning variation, as in Experiment 1.

3.3. Summary and discussion

Six-year-old children and adults were taught semi-artificial languages similar to those in Experiment 1, but where particle use was partially rather than fully predictable according to speaker (each speaker used their preferred particle on 75% of trials). Our analyses explored both learning of speaker identity conditioning, and regularization.

The ability to learn the speaker cues was again assessed via both a production task (day 1 and 4) and a 2AFC task (day 4). Both age groups were able to pick up on the speaker cues, although there were developmental and task differences. Adults showed significant learning and use of speaker-based conditioning in the production task on day 4, whereas children were at chance on both days. In contrast, both adults and children showed above chance performance in the 2AFC task (administered on day 4), showing sensitivity to the fact that the speakers had probabilistic preferences for different particles.

Why do we see this mismatch between production and acceptance/recognition behaviour for children? One possibility that we can rule out is failure to understand the task

of producing the language "for" one of the two speakers–the task here was identical to that in Experiment 1, where children were able to demonstrate the appropriate behaviour. More generally, it may be that children found the production task more taxing than the 2AFC task, so that only the latter is sufficiently sensitive to display the less robust learning in Experiment 2. In fact, there is precedent in the literature for dissociations between children's judgment of variation and their ability to produce it: In Hudson Kam and Newport (2009), children's performance in a judgment task reflected some sensitivity to determiner frequencies in the input, even though the use of the same determiners was highly systematized (i.e., regular) in a production task. More generally, there is evidence of greater learning in forced choice comprehension tasks than production tasks (Fisher, 2002; Gertner, Fisher, & Eisengart, 2006; Tomasello, 2000; Tomasello & Abbot-Smith, 2002) which has been interpreted in terms of production requiring more robust representations. It is also worth noting that adults showed increased use of speaker appropriate particles with longer exposure, and that even on day 4 they continued to undershoot the probabilities of their input, producing the majority particle around 62% of the time rather than 75%. If children's learning is generally slower than adults', it may be that they would eventually show evidence of conditioning in their production data if they were given additional training.

Regardless of the explanation of task differences, the results of the 2AFC test establish that both adults and 5-6 year-old children are sensitive to partial conditioning on speaker identity. As in Experiment 1, we also looked at whether participants' responses to the questionnaire showed awareness of the conditioning. Again, this was the case for more for adults than children. Interestingly, once "aware" participants were removed from the analyses, children's 2AFC responses continued to show sensitivity to speaker identity while adults' did not (although we acknowledged that this finding may reflect reduced power in the analyses of adults' performance, $n = 18$, which was less of a concern in the analyses of

children's performance, $n = 27$). We return to consider the role of awareness in the General Discussion. In sum, the results of this experiment extend the findings of Experiment 1: by age 5-6, children can track sociolinguistic variation conditioning on speaker identity under incidental conditions, even when that conditioning is (more realistically) probabilistic, and even though they may not be aware of the conditioning cues in question and have difficulty producing similar conditioning in their own output.

As in Experiment 1, we also looked for evidence of regularization in the production data. Again, we looked for both whole language regularization (boosting the frequency of one particle over the other) and lexical regularization (producing particular particles more frequently with particular lexical items). For whole language regularization, the pattern of results was very similar to Experiment 1: Children showed significant regularization while adults did not; regularization in children decreased from day 1 to day 4; and for children, there was greater regularization with novel nouns than old nouns. Comparing the number of significant regularizers in Experiment 2 relative to Experiment 1 showed that children's regularization was greater than in Experiment 1 on one of the days, suggesting that learning of the conditioning on speaker mitigated regularization to some extent. In the current experiment, there was also evidence of lexically based conditioning. Interestingly, age effects for this type of regularization were reversed, with only adults showing evidence of lexical conditioning. Recall that in Experiment 1 we did not see this type of regularization from adults. However, in that experiment, adults rapidly learnt the conditioning pattern. Thus, as for regularization in children, learning the conditioning on speaker works against the tendency to condition lexically for adults. Similarly, adults also showed more lexically based regularization on day 1 than day 4, so that lexical conditioning decreased with increased learning of the speaker identity cues.

We return to consider possible explanations of lexically based regularization, and why it is only seen in adults, in the General Discussion. For now, we point out that adults show a propensity to introducing this type of conditioning, which is not present at all in their input, but show no evidence of increasing the conditioning around speaker identity which is actually in their input: as mentioned above, they undershoot, rather than overshoot the speaker-based probabilities in their input.

For whole-language regularization, as in Experiment 1 we again see the predicted age difference (more regularization in children than adults), yet still much less regularization in children than seen in experiments involving fully artificial languages. For example, in Hudson Kam and Newport (2005), 71.3% of child participants showed strong (i.e., near absolute[11]) regularization, whereas in our experiment, this was only true for 17% of child participants by the final session. One possibility is that, even though children's production data did not reflect sensitivity to the speaker conditioning, learning of that conditioning nevertheless prevented strong regularization (cf. Hudson Kam, 2015, described in section 1.3). To explore this possibility, in Experiment 3, we exposed learners to an artificial language matched to that in Experiment 1 and 2, but where there was *no* conditioning of particle usage on speaker identity. Note that in this new artificial language, usage of particles is in free variation. This allows us to compare the results more directly with previous studies and to explore the factors affecting regularization given unconditioned variation.

4. Experiment 3

[11]Hudson Kam and Newport (2005) classified participants as regularizers if they produced the same form in all or all-but-one of their productions. The 17% we mention here were all participants who produced just one particle– none of our participants produced all-but-one regularization.

4.1. Methods

4.1.1. Participants

Thirty 5-6 year old children and 30 adults without any known hearing, language, or speech disorders participated in the experiment (Table 1). All participants completed the study on four consecutive days. Children were recruited from Year 1 and 2 of public primary schools in Coventry, Warwickshire, Greater London, Hertfordshire, and Scotland. Half of our participants were monolingual native English speakers and the other half spoke another language in addition to English. Adults were recruited through a university participant panel and received monetary compensation.

4.1.2. Stimuli, design and procedure

The stimuli and procedure were identical to Experiments 1 and 2. In Experiment 3, the consistency of particle usage in the learning input was manipulated such that particles were *maximally inconsistent:* both speakers used particle 1 50% of the time and particle 2 50% of the time. As in Experiments 1 and 2, this distribution of particles applied across the entire training set and also for each noun, such that particles 1 and 2 followed each training noun equal number of times.

Results

4.1.3. Production and 2AFC data

In Experiment 3, there was no correct or speaker-appropriate form for production or 2AFC performance, thus, these analyses were not carried out.

4.1.4. Regularization

4.1.4.1. Overall regularization

Figure 5 (first row, right panel) shows entropy scores for participants' productions. We first analysed whole-language entropy in a mixed ANOVA with age group (adults, children) as a between-subjects factor and day (day 1, day 4) as a within-subject factor. Mean whole-language entropy was significantly different relative to chance entropy (0.988, obtained via simulation), (mean = 0.71, 95% CIs = [0.63, 0.80]); there was also a significant main effect of age group ($p < .001$), a main effect of day ($p < .001$) and a significant interaction between age group and day ($p = .018$): The difference in mean entropy between day 1 and day 4 was significant both for children (mean difference = 0.33; 95% CIs = [0.16, 0.49]), and adults (mean difference = 0.08; 95% CIs = [0.006, 0.18]). Comparing each age group's regularization performance against chance showed that children's entropy was significantly below chance (0.988) on both days (day 1: 95% CIs = [0.21, 0.49]; day 4: 95% CIs = [0.54, 0.80]), whereas adults' entropy was below chance levels on day 1 (95% CIs = [0.75, 0.97]) and at chance on day 4 (95% CIs = [0.90, 0.989]). Note that, unlike Experiment 1, we do not see adults significantly increasing entropy on day 4, which only occurred when they were learning fully consistent speaker-based conditioning.

To see if there was any difference in regularization across old and novel nouns, we also subjected the data from old and novel nouns to an Age Group by Day by Noun Type mixed ANOVA. This analysis confirmed the main effects of age group and day and their interaction, but also found a main effect of noun type ($p = .019$) qualified by an interaction between noun type and age group ($p = .020$) (no other effects, ps > .372). Children showed significantly more regularization for novel nouns (mean entropy = 0.45) than old nouns (mean entropy = 0.54) (mean difference = -0.09; 95% CIs = [-0.14, -0.04]), but adults did not show a significant difference between old nouns (mean entropy = 0.87) and novel nouns (mean entropy = 0.87) (mean difference = 0.00; 95% CIs = [-0.05, 0.05]).

As in Experiments 1 and 2, we also used simulation to calculate the number of adult and child participants who reduced entropy to a greater extent than would be predicted by chance. Classified in this way, 26/30 children and 8/30 adults regularized on day 1 (a significant age difference: $\chi^2 = 21.99$, $p < .001$); 20/30 children and 4/30 adults regularized on day 4 (a significant age difference: $\chi^2 = 17.78$, $p < .001$). The following were absolute regularizers (i.e., only used one particle): 13/26 children and 3/8 adults on day 1, 5/20 children and 0/4 adults on day 4. Again, as seen in Experiment 1 and 2, substantial numbers of children produced data which is less variable than their input while *still* exhibiting some variability.

We again compared regularization across experiments by comparing the number of significant regularizers in each case. Compared with Experiment 1 (fully consistent language), more children regularized both on day 1 (13/28 in Experiment 1 vs. 26/30 in Experiment 3: $\chi^2 = 10.65$, $p = .001$) and day 4 (11/30 in Experiment 1 vs. 20/30 in Experiment 3; $\chi^2 = 5.41$, $p = .020$). For adults, there were more regularizers on day 1 (2/30 in Experiment 1 vs. 8/30 in Experiment 3; $\chi^2 = 4.32$, $p = .038$) but there were no differences in the number of adult regularizers on day 4 (0/30 in Experiment 1 vs. 4/30 in Experiment 3; $p = .112$, two-tailed Fisher-exact test). Compared with Experiment 2 (partially consistent language), there was no significant difference in the number of regularizers for either children or adults on either day (children day 1: 23/30 in Experiment 2 vs. 26/30 in Experiment 3: $\chi^2 = 1.00$, $p = .317$; children day 4: 18/30 versus 20/30; $\chi^2 = 0.29$, $p = .592$; adults day 1: 3/30 in Experiment 2 vs. 8/30 in Experiment 3, $\chi^2 = 2.78$, $p = .095$; adults day 4: 0/30 vs. 4/30; $p = .112$, two-tailed Fisher-exact test).

4.1.4.2. Lexical conditioning

Lexical MI was computed as in Experiments 1 and 2; results are shown in Figure 5 (second row, right panel) (recall that here high scores indicate *more* lexically based regularization). We first analysed whole-language MI in a mixed ANOVA with age group (adults, children) as a between-subjects factor and day (day 1, day 4) as a within-subjects factor. Overall, lexical MI was significantly different from chance MI (0.074, obtained via simulation) (mean MI = 0.20; 95% CIs = [0.16, 0.25]) and exhibited a significant main effect of age group ($p < .001$): Adults' productions had higher MI (mean MI = 0.30) (i.e., more lexically based regularization) than children's productions (mean MI = 0.10), and only adults showed MI which was significantly above chance (adults: 95% CIs = [0.25, 0.38]; children: 95% CIs = [0.065, 0.14]). There was no main effect of day ($p = .646$) and no significant interaction between age group and day ($p = .240$).

The analyses using the data from old and novel nouns (Age Group by Day by Noun Type mixed ANOVA) confirmed the main effect of age group and showed no effect of noun type ($p = .057$), no interaction with noun type (noun type by age group: $p = .864$; noun type by day: $p = .549$; noun type by age group by day: $p = .325$).

In addition, we also computed via simulation, for each participant, whether their MI score differed from that predicted by chance, given both the number of each of the two particles that they produced and the number of number of nouns that they produced. Classified in this way, 2/30 children versus 20/30 adults produced lexically conditioned output on day 1 (a significant age difference: $\chi^2 = 23.25$, $p < .001$); and 3/30 children versus 13/30 adults did so on day 4 ($\chi^2 = 8.52$, $p = .004$).

We also analyzed whether more participants lexically regularized in this experiment compared with those who lexically regularized in Experiment 1 and Experiment 2. Recall that in Experiment 1 there was no lexically based conditioning: There were more lexical conditioners in this experiment relative to Experiment 1 both on day 1 (0/30 in Experiment 1

vs. 20/30 in Experiment 3; $\chi^2 = 30.00$, $p < .001$) and day 4 (0/30 in Experiment 1 vs. 13/30 in Experiment 3; $\chi^2 = 16.60$, $p < .001$), which was not the case for children either on day 1 (0/28 in Experiment 1 vs. 2/30 in Experiment 3; $p = .492$, two-tailed Fisher-exact test) or day 4 (0/30 in Experiment 1 vs. 3/30 in Experiment 3: $p = .237$, two-tailed Fisher-exact test).

Compared with Experiment 2 (partially consistent language), the number of regularizers did not significantly differ for either group on either day (adults day 1: 15/30 in Experiment 2 vs. 20/30 in Experiment 3, $\chi^2 = 1.71$, $p = .190$; adults day 4: 9/30 in Experiment 2 vs. 13/30 in Experiment 3, $\chi^2 = 1.15$, $p = .284$; children day 1: 0/30 in Experiment 2 vs. 2/30 in Experiment 3, $p = .492$, two-tailed Fisher-exact test; children day 4: 3/30 in Experiment 2 vs. 1/30 in Experiment 3, $p = .612$, two-tailed Fisher-exact test).

4.1.4.3. Speaker-based conditioning

Although there was no relationship between speaker identity and particle usage in the input, it is possible that participants could nevertheless spontaneously introduce this type of regularity into their own productions, as some participants did when introducing lexical conditioning of variation. To explore this, we calculated an MI statistic equivalent to that used to explore lexical conditioning, by computing the entropy of particle usage for each of the two speakers separately, and subtracting the average of that entropy from total entropy:

$$MI_{speaker} = \textit{total entropy} - \textit{conditional entropy of particle given speaker}$$

The data are shown in Figure 8. For comparison purposes, we also computed the equivalent $MI_{speaker}$ scores for participants in Experiments 1 and 2 whereby participants were exposed to input that *did* feature speaker-based conditioning. Note that the analyses over $MI_{speaker}$ were not necessary for those experiments since they are essentially equivalent to the

analyses over particle production: In Experiment 1, both adults and children showed strong conditioning on speaker identity. In Experiment 2, this was only seen in adults and only on day 4, although they undershot the probabilities of the input. This pattern is also evident in Figure 8: In Experiment 1, children and adults both identified the speaker-based conditioning and produced output exhibiting significantly nonzero $MI_{speaker}$. In Experiment 2, adults (but not children) produced output with $MI_{speaker}$ values approximating those seen in their input, but only on day 4. In Experiment 3, $MI_{speaker}$ values are low across both days, indicating little evidence of this type of conditioning. For this experiment only, we ran mixed ANOVAs with equivalent structure to those used for the analyses over entropy and lexical MI.

For whole-language regularization around speaker identity, the ANOVA with age group (children, adults) as a between-subjects factor and day (day 1, day 4) as a within-subject factor revealed that the $MI_{speaker}$ did not differ from chance (0.0102, obtained via simulation), (mean = 0.03; 95% CIs = [0.0101, 0.05]), and that there was no significant effect of age group (p = .215), day (p = .863), or age by day interaction (p = .184). The mixed ANOVA over the data from old and novel nouns confirmed this pattern of results and further demonstrated that there was no main effect of noun type (p = .846) or interactions with noun type (noun type by age group: p = .391; noun type by day: p = .190; noun type by age group by day: p = .521).

We also computed via simulation, for each participant, whether their $MI_{speaker}$ values differed from those predicted by chance, given both the number of each of the two particles that they produced and the number of productions for each participant: on day 1, 5/30 adults and 0/28 children (a significant age difference, p = .053, two-tailed Fisher-exact test) showed significant conditioning; on day 4, 2/30 adults and 1/30 children showed significant conditioning (p = 1.000, two-tailed Fisher-exact test).

For comparison purposes, we also went back and computed the number of participants in the previous two experiments who were, according to $MI_{speaker}$, conditioning their particle choice on speaker identity[12]. In Experiment 1, it was 19/28 children and 28/30 adults on day 1 and 21/30 children and 30/30 adults on day 4. In Experiment 2, it was 2/30 children and 8/30 adults on day 1, and 5/30 children and 13/30 adults on day 4. All comparisons between Experiment 1 and Experiment 3 were significant (*ps* < .001), while for the comparison between Experiment 2 and Experiment 3, only the comparison for adults on the final day was significant ($\chi^2 = 10.76$, $p = .001$; all other comparisons *p*s > .195). This corroborates the previous analyses and reflects (i) the significant learning of the conditioning on both days for both groups in Experiment 1, and (ii) the fact that in Experiment 2, only adults on day 4 showed significant conditioning.

4.1.4.4. Summary of regularization data: Experiment 3

Children showed greater regularization (i.e., reduction of entropy) in particle usage across the language relative to adults, and their regularization decreased from day 1 to day 4, as exposure to the language increased. In children only, there was evidence of greater regularization with novel nouns relative to old nouns. Children showed greater regularization in Experiment 3 compared to Experiment 1 (where children did learn the conditioning factor) but not compared to Experiment 2 (where their production data suggests they did not). Adults' tendency to regularize decreased from day 1 to day 4 and was greater than chance on day 1, but not on day 4. There were also more adult regularizers in Experiment 3 than in

[12] Note that we use the same somewhat conservative measure as we use for identifying regularizers, i.e., that there is only a 5% chance that their MI could would be this high due to chance.

Experiment 1. These results suggest that adults may have a stronger tendency to regularize in the absence of conditioning factors that explain the variation seen in their input.

Turning to lexically based conditioning, as in Experiment 2, we saw this in adults but not significantly in children. Unlike Experiment 2, adults did not reduce their level of lexical conditioning on day 4. There also appeared to be more lexically based conditioning than in Experiment 2, however, comparison of the number of significant lexical-regularizers did not reveal significant differences.

4.2. Summary and discussion

Six-year-old children and adults were taught a semi-artificial language equivalent to those used in Experiment 1 and 2, but where the usage of particles was maximally inconsistent, with each of the two speakers producing each particle equally often with each noun. Our analyses focused on regularization within the production data. We looked for both whole language regularization (boosting the frequency of one particle over the other) and lexical conditioning (producing particular particles more frequently with particular lexical items).

The pattern of results was very similar to that seen in the previous experiments. For whole language regularization, we found that (i) children showed greater regularization relative to adults; (ii) regularization decreased from day 1 to day 4 in both age groups; and (iii) at least for children, there was greater regularization with novel nouns than old nouns. As in Experiment 2, adults introduced more lexical conditioning than children. Unlike Experiment 2, their lexical conditioning was relatively consistent across days, presumably due to the fact that, in this experiment, they were not learning alternative conditioning from the input. In general, as for Experiment 2, we saw more regularization (whole-language and lexical conditioning) compared to Experiment 1. This suggests, again, that regularization is

reduced when participants show strong learning of alternative patterns in the input. Overall, these findings corroborate those from the previous experiments. However, note that this experiment goes beyond replication: here we can be sure that differences in the extent of regularization across age groups and days are *not* due to learning of conditioning patterns which are actually present in the input. Thus, we show that, even in the absence of conditioning factors, children's regularization still decreases with exposure. This suggests that increased exposure makes them more willing or more able to replicate unpredictable variation present in their input. Similarly, the greater regularization seen in children cannot not simply be due to the weaker learning of conditioning factors present in their input: rather, children have a stronger-but not absolute-bias against producing inconsistent variation. The fact that children regularize more with novel nouns and less with familiar nouns can be interpreted as indicating that learner biases against unpredictable variation are modulated by the evidence for variability they in the input, i.e. children are more willing to produce variation with those items with which it occurred in their input.

We also considered a final type of regularization, namely, speaker-based conditioning. Participants in Experiment 3 received input where both speakers used each particle equally; however, they could themselves produce the particles at different frequencies for each speaker (introducing speaker-based conditioning), in a similar way to which they introduced lexically conditioning. Overall, the analyses suggested no significant evidence of this type of conditioning, with only a handful of participants categorized as introducing significant speaker-based conditioning. This is in line with the fact that we saw no tendency to increase conditioning of particle usage on speaker in Experiment 2, which would again be a type of regularization. The fact that we see a clear bias to regularize around lexical items but no corresponding bias to regularize around speaker is also in line with the

findings of a recent iterated learning experiment suggesting a relative lack of sensitivity to speaker identity cues (Smith et al., 2017).

5. General Discussion

Languages exhibit variation which may be conditioned or partially conditioned on sociolinguistic cues such as social context or aspects of the speaker's group identity. Previous work suggests that children's speech sometimes matches these conditioning factors; however, it is not clear whether this demonstrates that they are tracking the underlying statistics or whether they have been exposed to differential input (Smith et al., 2007, 2013). We introduced an experimental methodology for exploring children's ability to acquire deterministic or probabilistic sociolinguistic variation. Using a semi-artificial language learning task, we probed the ability of children and adults to learn that the use of variants is conditioned on speaker identity. In Experiment 1 (deterministic conditioning), the two particles were equally frequent in the learner's input, but their use was entirely predicted by speaker identity (one speaker always used particle 1 and the other speaker always used particle 2). In Experiment 2 (probabilistic conditioning), both speakers used both particles, but had different probabilistic preferences (speaker 1 used particle 1 75% of the time, speaker 2 used particle 2 75% of the time). In Experiment 3 (unconditioned or "free" variation), both speakers used both particles equally frequently. Participants undertook production tasks at the end of day 1 and day 4 in which they were prompted to complete sentences as produced by one of the two speakers. At the end of the last session, participants also took a 2AFC test recording whether they chose the option with the majority form for that speaker. Although learning of the speaker identity conditioning cue was most robust in the deterministic language, and stronger for adults than children, we saw evidence for learning of the conditioning cue even for children in the partially consistent, probabilistically-conditioned

language. The results thus establish that both adults and 5- to 6-year-olds are sensitive to speaker identity cues which condition linguistic variation. In addition, we saw evidence of regularization in all three experiments, involving either boosting of the frequency of one particle over the other (seen more in children than adults), or conditioning of particle use on lexical items, seen more in adults than children. We now discuss the implications of these findings for the literatures on developmental sociolinguistics and regularization of unpredictable variation.

5.1. Sociolinguistic development

These results establish that 5- to 6-year-olds are able to pick up on the fact that different speakers use different variants, even when that relationship is probabilistic. This basic ability is key to learning any sociolinguistic variation which is cued by some aspect of group identity, and the result is in line with various demonstrations in the literature where children's production of sociolinguistic variation is predictable from factors such as gender and socioeconomic status (Díaz-Campos, 2005; Foulkes et al., 2005; Habib, in press). Problematically, the suggestive patterns of sociolinguistic conditioning seen in the naturalistic data are also consistent with the possibility that children's socially conditioned variant use reflects systematically different input to children of different social groups (e.g., the linguistic input girls hear might be different from that received by boys; Foulkes et al., 2005). Even for style shifting, where the same child appears to show different usage of variants in different social contexts, it can be difficult to pull apart actual sociolinguistic sensitivity (i.e., knowledge that particular forms are more appropriate in particular contexts) from a bias to align with parental frequencies within the immediate environment, particularly given that data are often taken from parent-child interactions (e.g., Smith et al., 2007, 2013). Importantly, in the current experiments, our knowledge of the input allows us to be certain

that participants are actually acquiring a pattern of conditioned variation evidenced in their input, i.e., they are capable of learning to condition linguistic variation on social cues.

We conducted post-experiment interviews which included questions designed to elicit any explicit knowledge of the relationships between speaker identity and particle choice. In both Experiments 1 and 2, both children and adults were able to verbalize this relationship. Unsurprisingly, this occurred more in adults, and more given consistent (Experiment 1) than partially consistent input (Experiment 2). However, follow-up analyses suggested that, at least in children, learning of the patterns did not depend upon being able to verbalize the relationships. This is in line with some previous research (with adults) using artificial learning methods which found significant learning of connections between meaning (the animacy of a noun) and form (the usage of two novel particles) in participants who were unable to report the relationship (Leung & Williams, 2011; Williams, 2005). In general, however, the relationship between implicit and explicit learning in artificial language learning paradigms is not well-understood. It may be that awareness is a product of successful implicit learning, rather than reflecting involvement of a separate, explicit learning mechanism. It is also unclear whether a failure to verbalize the patterns, as commonly seen in children, implies truly implicit learning, or just a failure of articulacy. Nevertheless, our results do suggest that children can pick up on sociolinguistic conditioning in the absence of explicit awareness. This is interesting in the context of the sociolinguistic literature where adult users are often highly aware of conditioning on factors such as gender, and the usage of the variants may be high emotionally charged. Our results suggests that this type of sensitivity is not a prerequisite for sociolinguistic learning.

5.2. Regularization

Previous experiments suggest that learners do not always veridically acquire the variation that is present in their input but may regularize, either by reducing or eliminating the variation, or by conditioning it on some aspect of the linguistic context (Hudson Kam & Newport, 2005, 2009; Smith & Wonnacott, 2010). In the current experiments, we used information theoretic techniques to identity and quantify different types of regularization. We found that regularization occurred across all three experiments, i.e., both when variation in the input was fully or partially conditioned on the speaker, and when it was completely unconditioned. The results were remarkably consistent across experiments, suggesting that regularization tends to occur both when variation is completely unpredictable (Experiment 3) and when it is actually conditioned (Experiments 1 & 2), at least in the early stages of learning. In the latter case, learner-imposed regularization/conditioning diminishes as the "real" underlying conditioning cues are identified.

One set of analyses looked at the tendency to regularize by increasing the usage of one variant across the language (i.e., reducing total entropy). Key findings were that: (i) in line with previous literature (e.g., Hudson Kam & Newport, 2005, 2009), children showed more regularization than adults (with just a handful of adults regularizing in this way, and regularization not significant at the group level for adults anywhere but on day 1 of experiment 3); (ii) children showed weaker tendencies to regularize than seen in some previous studies, and by day 4 the majority of them (approximately 83% in each experiment) produced variable output (i.e., produced both particles on at least some trials); (iii) both adults and children showed less regularization with more exposure; and (iv) children regularized more with familiar than unfamiliar (i.e., novel) nouns.

Why do children show more regularization than adults? One explanation in the literature is that children's greater tendency to regularize is due to their more limited memory resources (Hudson Kam & Newport, 2009). This is consistent with the fact that our learners

decreased their regularization with increased exposure (more input is needed to overcome capacity limitations). It also fits with the fact that children in these experiments show less regularization than in previous experiments by Hudson Kam and Newport (2005, 2009), since various aspects of our design reduce the burden on memory. Specifically: we use familiar (English) nouns rather than novel nouns; the set of exposure sentences is smaller and simpler than typically used in previous experiments (shorter sentences using fewer words); each sentence is repeated several times; we include exposure just prior to testing (whereas Hudson Kam and Newport generally had the test in a separate session). If memorization of the input is key, these factors could potentially reduce regularization. On the other hand, there is also some evidence which speaks against a straightforward memory-based account of regularization. Perfors (2012) presented a series of experiments which demonstrate that putting adults under "load" (in a variety of ways) during the learning process has no effect on later regularization in a production task. In addition, she found no relationship between participants' performance in a battery of memory tests and the extent of their regularization (similarly, exploratory analyses with our own data set also found no evidence of correlation between the tests of working memory which we conducted and our measures of regularization – see footnote 4). Perfors (2012) therefore concludes that memory constraints during learning are not responsible for regularization, and thus, are unlikely to be the cause of age differences in regularization.

An alternative explanation of age differences is that children are somehow more constrained at the point of retrieval rather than encoding, i.e., during production rather than learning, and that retrieval difficulties are also reduced with more exposure (accounting for the consistent effects of day in our data). This fits with the results of an adult artificial language study by Hudson Kam and Chang (2009), which found that adults regularized less when provided with a verbal reminder of the relevant verb and noun prior to each test trial, or

when presented with flashcards on which all on the verbs, nouns, and particles in the language were written.

Note that the idea that retrieval difficulties increase regularization is also consistent with the weaker regularization seen in these studies than in previous artificial language studies by Hudson Kam and Newport (2005, 2009): as mentioned above, there are several simplifying factors in this experiment, but one which is particularly relevant to retrieval is the fact that we used semi-artificial languages, where the nouns are familiar, and thus presumably more easily retrieved compared with the novel vocabulary used in their fully artificial languages[13]. A retrieval explanation would also fit with the fact that, consistently across the three experiments, children showed more regularization for untrained than trained nouns (as seen in Wonnacott & Newport, 2005, although this was for adult learners with fully artificial languages). This is difficult to explain in terms of memorization of the input utterances at encoding: if participants were simply memorizing the training sentences, one would have expected to see the difference between trained and untrained nouns in adults, who are presumably better able to memorize than children. However, in terms of retrieval, children (but not adults) would be expected to have more difficulty retrieving the words which have not occurred in their input (even though they are familiar from English), making untrained nouns harder for them to access and thus potentially explaining why they showed more regularization with novel nouns in the experiments. Whether it is true that retrieval difficulties exacerbate regularization remains an open question. Regardless, it is important to realize that there must necessarily be an inherent bias for regularization which is present in

[13]Of course, there are other differences between the experiments, most notably the fact that there is no "dominant" form in the input. However, note that Wonnacott (2011) taught children a similar semi-artificial language where one particle *was* more frequent than the other and also found a similar pattern of reduced regularization

the learner in order for this behaviour to emerge when the system is under pressure. Perfors (2012) makes a similar point (although her focus was on memory load during learning, rather than at the point of retrieval). She produced a computational model which learned distributions of linguistic variants and simulated putting the system under memory load given various priors for regularization. The results suggested that regularization can occur under load, but only when learners have an inherent prior bias favouring regularization (Perfors, 2012).

Other explanations of children's greater regularization have also been proposed in the literature. Ramscar and colleagues (Ramscar & Gittcho, 2007; Ramscar & Yarlett, 2007) suggest that regularization naturally emerges in a model of morphology learning, where items are reactivated associatively in memory as a function of experience, and the form to be produced is determined by brute strength competition such that the strongest activation gets produced. In contrast, probability matching requires an additional process for selecting between responses which relies on a type of goal relevant conflict resolution which is more mature in adults. Rische and Komarova (2016) present a related learning model which both increases the probability of a particle whenever it is encountered (positive evidence) and decreases the probabilities of any alternative particles which were not encountered (negative evidence). They demonstrate that regularization may emerge when participants are more biased to respond to positive than negative evidence, and suggest that this may underpin the difference between children and adults. This model can capture the regularization data in Hudson Kam and Newport (2009); however, one aspect of the current data which we suspect might not be captured by their model is the fact that children's regularization decreased with increased exposure.

In our study, we statistically contrasted, for the first time, children's and adults' tendency to condition particle use on linguistic context, specifically on lexical items. Adults

introduced lexical conditioning when social conditioning was nondeterministic (Experiment 2) or absent (Experiment 3). Only a handful of children showed lexical conditioning in any experiment, and lexical conditioning was not significant at the group level for children; thus, we demonstrated that children are less likely to condition their usage of particles around lexical items than adults, corroborating an observation from Hudson Kam and Newport (2009). The demonstration that adults show lexically-based conditioning fits with the findings of the iterated artificial language learning experiment in Smith and Wonnacott (2010), where 6 of 10 chains of adult learners were found to transform an initially inconsistent language into one where particle usage was fully conditioned on the noun (3 of the remaining 4 chains ended up with "absolutely" regular usage of a single plural marker, and 1 chain exhibited variation which was not significantly lexically conditioned; in other words, lexical conditioning was the mode outcome of iterated learning, and the most common type of regularization outcome).

Where does this adult bias for lexicality come from and why is it not much seen in children? Previous experiments have established that both adults and children are *capable* of learning lexical conditioning when there is sufficient evidence of lexicality in the input (Wonnacott et al., 2008; Wonnacott 2011). One possible mechanism for lexical conditioning that we can rule out is that of immediate self-priming, since consecutive items in the production tests did not include the same noun. In addition, if priming were the explanation, it is unclear why we see so much less of this type of regularization in children. It is also worth noting that, although they are capable of learning speaker-based conditioning, neither adults nor children spontaneous condition on speaker identity in Experiment 3, or increase this conditioning in Experiment 2. This suggests that all types of conditioning are not a priori equally likely (see also Smith et al., 2017). Other types of conditioning have been reported in previous experiments, such as conditioning on sentence position, subject versus object, in

Hudson Kam (2016). This syntactic conditioning was not possible in our study: there was a single noun in each sentence and the same verb repeated across sentences, that is, participants could not condition on anything other than the noun. An interesting question which we address in our ongoing research program, is which types of conditioning contexts are more relevant for learners of different ages. Returning to the lexically based conditioning seen in the current experiment, one possible explanation of the fact that this is only seen in adults is that their wider experience of natural language has taught them that lexically based regularity (but not speaker-based regularity) is highly common and thus they seek to impose this on new input. Another is that the structure of the input is relevant, specifically the number of speakers versus lexical items (two vs. eight) means that they have more evidence of variability at the speaker level than at the lexical level. *Where* the lexical bias comes from, and *why* it is significantly weaker in children, remains to be seen. However, the current data raises the possibility that adults might play a larger role than children in introducing this particular type of regularity into morphological systems within natural languages.

Altogether, the results of the current experiment confirm that learners have a bias against reproducing fully inconsistent linguistic variation. Computational research, and results from iterated learning experiments, suggest that even weak learner biases may play a role in shaping human language once we allow for the magnifying effects of transmission across learners (Reali & Griffiths, 2009; Smith & Wonnacott, 2010); factors at play during language use (e.g. pragmatic reasoning, lexical or structural priming) might also interact with biases in learning to contribute to the types of variability we see in natural language (Fehér, Wonnacott & Smith, 2016; Perfors, 2016; Smith et al., 2017). Thus, the learning biases which we see in this and previous experiments may underpin the lack of free variation across the world's languages. However, the real-world finding that some individual children show strong regularization of inconsistent variation present in their input despite long-term

exposure to that variation (Singleton & Newport, 2004) is somewhat surprising given our finding that, with time, children can come to acquire unpredictably variable linguistic systems. One possibility is that the structure of the input received by such children is qualitatively different from the structure of the input in the current experiments. One possibility (Hudson Kam, 2000) is that Simon's input appears to deviate most from his input when his parents' input deviated most from each other. It is possible that children do not reproduce variation unless it occurs consistently across speakers (see Fehér, Kirby, & Smith, 2014, for an experimental study showing similar results). Another (admittedly vague) possibility is that the specifics of the type of linguistic variation is important and that we would see more elimination of inconsistent variation if the variants and the circumstances in which they occurred matched the naturally occurring circumstances more closely.

5.3. Limitations and future directions

The results of the current experiment establish that, by age 5-6, children can track variation conditioned on speaker identity, and that this ability continues and strengthens into adulthood. We chose to work with children aged 5 to 6 given that previous literature established their ability to follow procedures in these type of artificial language learning paradigms (Hudson Kam & Newport 2005, 2009; Hudson Kam, 2015; Wonnacott, 2011); working with the same age group also allows us to draw direct comparisons with those previous results. However, to fully establish the developmental trajectory of sociolinguistic learning, it will be important to develop paradigms for use with younger children.

This same paradigm could be used to investigate children's ability to learn different types of conditioning cue as well as different types of variation. For example, although the current work establishes that children can track the linguistic usage of different speakers, this is not sufficient (although we believe it is necessary) to generalize over the usage of different

groups. An ongoing experiment uses the current paradigm to explore whether children generalize over speaker gender by expanding the input to include multiple speakers of each gender, and new speakers at test. Similar questions can be asked about the types of linguistic cues which children of different ages can or cannot track; our ongoing work explores learners' sensitivity to different syntactic and semantic cues (cf. Hudson Kam, 2015). Although more methodologically challenging, it would also be interesting to expand these same techniques to explore the acquisition of other types of sociolinguistic conditioning, in particular those which do not rest on tracking speaker identity, but rather on different situational contexts (i.e., style shifting).

In addition to looking at conditioning on a wider range of cues, it would also be interesting to look at different types of variation. Here, we looked at lexical variation (in the post-nominal particle) which is likely to be particularly salient. However, other types of variation could be explored, for example phonological or syntactic variation (c.f. Wonnacott & Newport, 2005; Wonnacott et al., 2008). As reviewed in section 1.2, different types of sociolinguistic variation seem to exhibit rather different learning trajectories. It seems likely that difficulty of acquisition will be a function of the complexity and salience of both the conditioning factors and the type of variants in question. The current paradigm provides a methodology for pulling these factors apart.

In order to understand the mechanisms driving regularization, an important direction is to establish whether learning of a semi-artificial language does indeed lead to less regularization compared with learning of a precisely-matched but fully artificial language. We are currently running this experiment and hope that the results will speak to the extent to which retrieval plays a role in regularization as well as the different biases of adults and children (Wonnacott et al., in preparation). We are also working to understand the extent to which our results can be captured by the different models that have been proposed in the

literature (Smith et al., 2017; Ramscar & Yarlett, 2007; Rische & Komarova, 2016). Future work should also seek to establish the factors which lead to lasting regularization in the face of stably unpredictable input, as seen in the child Simon.

5.4. Conclusion

We have demonstrated that both 6-year-olds and adults can learn linguistic variation which is deterministically or probabilistically conditioned on speaker identity. This basic ability must underpin acquisition of any sociolinguistic variation in natural languages, wherever that conditioning is cued by some aspect of group identity. In line with previous experiments, we also saw evidence of regularization. We demonstrated, for the first time, that different types of regularization (whole-language reduction in variability vs. lexical conditioning) occur in children and adults. Our experimental paradigm provides a methodology for continuing to explore both how learners of different ages acquire different types of linguistic variation, and the phenomena of regularization.

Acknowledgements

This work was supported by the Economic and Social Research Council (grant number ES/K006339) held by the second and last author. We would like to thank the teachers and pupils of the participating schools in Warwickshire, East Lothian and London for their help and cooperation. Thanks also to Nicola Davies-Jenkins, Julie Joubert, and Gill Kinchen for assisting with data collection.

References

Alloway, T. P., Gathercole, S. E, Kirkwood, H. J., & Elliott, J. E. (2008). Evaluating the validity of the Automated Working Memory Assessment. *Educational Psychology, 7*, 725-734.

Austin, A. C., Newport, E. L., & Wonnacott, E. (2006). Predictable versus unpredictable variation: Regularization in adult and child learners. *Paper presented at the Boston University Conference on Child Language Development*, Boston, MA.

Baayen, R.H., Davidson, D.J. & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59 (4)*, 390-412. doi: 10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(1)*, 1-48. doi:10.18637/jss.v067.i01.

Braine, M. D. (1963). On learning the grammatical order of words. Psychological Review, 70(4), 323.

Cameron, D. (2005). Language, gender, and sexuality: Current issues and new directions. *Applied linguistics, 26(4)*, 482-502. doi: 10.1093/applin/ami027

Chambers, K. E., Onishi, K. H., & Fisher, C. (2010). A vowel is a vowel: Generalizing newly learned phonotactic constraints to new contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(3)*, 821-828. http://dx.doi.org/10.1037/a0018991

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition, 139,* 71-82. doi:10.1016/j.cognition.2015.02.007

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition, 122(3),* 306-329. doi:10.1016/j.cognition.2011.10.017

Díaz-Campos, M. (2005). "The emergence of adult-like command of sociolinguistic variables: A study of consonant weakening in Spanish-speaking children." In D. Eddington (Ed.), *Studies in the Acquisition of the Hispanic Languages: Papers from the 6th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*. Somerville, MA: Cascadilla, 56-65.

Eckert, P. (1997). Age as a sociolinguistic variable. In F. Coulmas (Ed.), *Handbook of sociolinguistics* (pp. 151-167). Oxford: Blackwell.

Eckert, P., & McConnell-Ginet, S. (1999). New generalizations and explanations in language and gender research. *Language in Society, 28,*185-201.

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. C. M. (2015). The relationship between artificial and second language learning. *Cognitive Science, 40*, 822-847. doi: 10.1111/cogs.12257

Fehér, O., Kirby, S., & Smith, K. (2014). Social influences on the regularization of unpredictable variation. In P. Bello, M. Guarini, M. McShane & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2187-2191). Austin, TX: Cognitive Science Society.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language, 91*, 158-180. http://dx.doi.org/10.1016/j.jml.2016.06.002

Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: a reply to Tomasello (2000). *Cognition, 82(3),* 259-278. doi:10.1016/S0010-0277(01)00159-7

Foulkes, P., Docherty, G. J., & Watt, D. J. (2001). The emergence of structured variation. *University of Pennsylvania Working Papers in Linguistics, 73(3)*, 67-84.

Foulkes, P., Docherty, G. J., & Watt, D. J. (2005). Phonological variation in child-directed speech. *Language, 81 (1)*, 177-206.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language, 39(2)*, 218-245. doi: 10.1006/jmla.1998.2569

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32(02)*, 249-268. doi: http://dx.doi.org/10.1017/S0305000904006786

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science, 17(8)*, 684-691. doi: 10.1111/j.1467-9280.2006.01767.x

Habib, R. (in press). Parents and their children's variable language: Is it acquisition or more? *Journal of Child Language*.

Hama, M. & Leow, R. P. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition, 32*, 465-491. http://dx.doi.org/10.1017/S0272263110000045

Hoffman, M., & Walker, J. (2010). Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change, 22*, 37-67. http://dx.doi.org/10.1017/S0954394509990238

Hudson, C. (2000). Variation: When can it be learned and when does it lead to language change? Paper presented at the meeting of the Society for Pidgin and Creole Linguistics, Chicago, IL.

Hudson Kam, C. L. (2015). The impact of conditioning variables on the acquisition of variation in adult and child learners. *Language, 91 (4)*, 906-937. doi: 10.1353/lan.2015.0051

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. http://dx.doi.org/10.1037/a0015097

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*, 151-195. doi: 10.1080/15475441.2005.9684215

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology, 59*, 30-66. doi: 10.1016/j.cogpsych.2009.01.001

Jaeger, T.F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446. doi: 10.1016/j.jml.2007.11.007

Knack, R. (1991). Ethnic boundaries in linguistic variation. In P. Eckert (ed.), *New ways of analyzing sound change*. New York: Academic Press. 251-272.

Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change, 1(3)*, 199-244. http://dx.doi.org/10.1017/S0954394500000168

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44(4)*, 978-990. doi: 10.3758/s13428-012-0210-4

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

Labov, W. (1964). Phonological correlates of social stratification. *American Anthropologist, 66 (2),* 164-176.

Labov, W (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, W. (1989). The child as linguistic historian. *Language Variation and Change, 1*, 85-97.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change, 2(2),* 205-254. http://dx.doi.org/10.1017/S0954394500000338

Labov, W. (2001). *Principles of Linguistic Change: Social Factors*. Oxford: Blackwell.

Leung, J. H., & Williams, J. N. (2011). The implicit learning of mappings between forms and contextually derived meanings. *Studies in Second Language Acquisition, 33(1),* 33-55. http://dx.doi.org/10.1017/S0272263110000525

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: studies with artificial lexicons. *Journal of Experimental Psychology: General, 132(2),* 202-227. http://dx.doi.org/10.1037/0096-3445.132.2.202

Mendoza-Denton, N. (2008). *Homegirls: Language and cultural practice among Latina youth gangs*. Malden, MA: Blackwell Publishing.

Miller, K. & Schmitt, C. (2012) Variable Input and the Acquisition of Plural Morphology. *Language Acquisition: A Journal of Language Development, 19 (3),* 223-261.

Neu, H. (1980). Ranking of constraints on /t,d/ deletion in American English: A statistical analysis. In W. Labov (ed.), *Locating language in time and space*. New York: Academic Press.

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

Newport, E. L. (1999). Reduced input in the acquisition of signed languages: Contributions to the study of creolization. In M. Degraff (Ed.), *Creolization, diachrony, and language acquisition* (pp. 161-718). Cambridge, MA: MIT Press.

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development, 12(2),* 138-155. http://dx.doi.org/10.1080/15475441.2015.1052449

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language, 67(4),* 486-506. doi: 10.1016/j.jml.2012.07.009

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06*, University of Cambridge, Cambridge.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59,* 413-425. doi: 10.1016/j.jml.2008.02.002

R Core Team. (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. http://www.R-project.org/

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Science, 11(7),* 274-279. doi:10.1016/j.tics.2007.05.007

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science, 31 (6),* 927-960. doi: 10.1080/03640210701703576

Rische, J. L., & Komarova, N. L. (2016). Regularization of languages by adults and children: A mathematical framework. *Cognitive Psychology*, *84*, 1-30. doi:10.1016/j.cogpsych.2015.10.001

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition, 111(3)*, 317-328. doi:10.1016/j.cognition.2009.02.012

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology, 66*, 30-54. doi: 10.1016/j.cogpsych.2012.09.001

Reid, E. (1978). Social and stylistic variation in the speech of children: some evidence from Edinburgh. In P. Trudgill (ed.), *Sociolinguistic patterns in British English*, 158–171. London: Edward Arnold.

Roberts, J. (1994). *Acquisition of variable rules: (-t,d) deletion and (ing) production in preschool children*. Ph.D. dissertation, University of Pennsylvania.

Roberts, J. (1997). Acquisition of variable rules: A study of (-t, -d) deletion in preschool children. *Journal of Child Language, 24*, 351–372.

Romaine, S. (1978). Postvocalic /r/ in Scottish English: Sound change in progress? In P. Trudgill (ed.), *Sociolinguistic Patterns in British English* (pp. 144-157). Baltimore, MD: University Park Press.

Romaine, S. (1984). *The language of children and adolescents*. Oxford: Blackwell.

Ross, D. S., & Newport, E. L. (1996). The development of language from non-native linguistic input. In A. Stringfellow, D. Cahana-Amitay, E. Hughes, & A. Zukowski (Eds.) *Proceedings of the 20th Annual Boston University Conference on Language Development* (pp. 634–645). Boston: Cascadilla Press

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

Shin, N. L. (2016). Acquiring constraints on morphosyntactic variation: Children's Spanish subject pronoun expression. *Journal of Child Language 43(4),* 914-947. doi: http://dx.doi.org/10.1017/S0305000915000380

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from impoverished input. *Cognitive Psychology, 49,* 370-407. doi: 10.1016/j.cogpsych.2004.05.001

Smith, J., Durham, M., & Fortune, L. (2007). "Mam, ma troosers is fa'in doon!" Community, caregiver and child in the acquisition of variation in Scottish dialect. *Language Variation and Change, 19 (1),* 63-99. doi: http://dx.doi.org/10.1017/S0954394507070044

Smith, J., Durham, M., & Fortune, L. (2009). Universal and dialect-specific pathways of acquisition: Caregivers, children, and t/d deletion. *Language Variation and Change, 21(01),* 69-95. doi: http://dx.doi.org/10.1017/S0954394509000039

Smith, J., Durham, M., & Richards, H. (2013). The social and linguistic in the acquisition of sociolinguistic norms: caregivers, children, and variation. *Linguistics, 51(2),* 285-324. doi: 10.1515/ling-2013-0012

Smith K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott E. (2017). Language learning, language use, and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B, 372(1711),* 20160051. http://dx.doi.org/10.1098/rstb.2016.0051

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition, 116(3),* 444-449. doi:10.1016/j.cognition.2010.06.004

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568. doi:10.1016/j.cognition.2007.06.010

Sorace, A. (2000). Gradients in auxiliary selection with intransitive verbs. *Language, 76(4),* 859-890.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74(3),* 209-253. doi:10.1016/S0010-0277(99)00069-4

Tomasello, M. & Abbot-Smith, K. (2002). A tale of two theories: Response to Fisher. *Cognition, 83(3),* 207-214. doi: 10.1016/S0010-0277(01)00172-X

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* (3rd edition). Waltham, MA: Academic Press.

Williams, J. N. (2004). Implicit learning of form-meaning connections. In J. Williams & B. VanPatten & S. Rott & M. Overstreet (Eds.), *Form-meaning connections in second language acquisition* (pp. 203-218). Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition, 27*, 269-304.

Wolfram, W. (1969). *A Sociolinguistic Description of Detroit Negro Speech*. Washington DC: Center for Applied Linguistics.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language, 65*, 1-14. doi: 10.1016/j.jml.2011.03.001

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5year olds. *Journal of Memory and Language, 66(3),* 458-478. doi:10.1016/j.jml.2011.11.004

Wonnacott, E., & Newport, E. (2005). Novelty and regularization: The effect of novel instances on rule formation. In A. Brugos, M. Clark-Cotton, & S. Ha (Eds.), *BUCLD 29: Proceedings of the 29th annual Boston University conference on language development*. Boston, MA: Cascadilla Press.

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. doi: 10.1016/j.cogpsych.2007.04.002

Wonnacott, E., Samara, A., Saul, J., & Smith, K. (in preparation). When to be regular: Factors that impact on regularization in artificial language learning studies with children.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18(5),* 414-420. doi: 10.1111/j.1467-9280.2007.01915.x

Appendix A

Post-Experiment Questionnaire

(1) What do you think “glim” means?

(2) What do you think X and Y mean (where X and Y are the two particles)?

(3) How did you decide when to use X and when to use Y?

(4) Did you notice any patterns in how X and Y were used when you were repeating the sentences for Henry and Katie?

(5) How early in the experiment did you notice these patterns?

(6) Adults Only: Did you consider any other possible patterns along the way that turned out to be incorrect?

(7) Adults Only: Did you focus more on learning which of the new words went with each individual picture (e.g., which word went with dog), or on which of the words went with each voice (e.g., which new word the female voice used)?

Appendix B

Analyses of production and 2AFC task performance in Experiment 1 excluding "aware" child participants.

Choice of particle

Figure B9 shows the proportion of speaker appropriate particles produced by "unaware" children in experiment 1 (n = 18) as a function of noun type and day. The model predicting production of the correct particle showed a significant intercept (b = 2.64, SE = 0.51, z = 5.13, p < .001), suggesting that performance was greater than chance (50%), no significant effect of day (b = -0.35, SE = 0.26, z = -1.32, p = .183), noun type (b = -0.23, SE = 0.25, z = -0.94, p = .344), and no day by noun type interaction (b = 0.28, SE = 0.50, z = 0.57, p = .569). These analyses suggest that the fully consistent speaker identity cues were picked up by children who could not verbally report the association between speaker identity and particle choice.

2AFC data

"Unaware" children's 2AFC task performance in Experiment 1 across old and novel nouns is shown in Figure B10. The model predicting particle choice showed a significant intercept (b = 3.47, SE = 1.21, z = 2.88, p = .004), suggesting that performance was greater than chance, and no significant main effect of noun type (b = 0.10, SE = 1.13, z = 0.09, p = .929). Thus, "unaware" children made speaker-appropriate particle choices more frequently than expected by chance in response to both noun types.

Appendix C

Analyses of production and 2AFC task performance in Experiment 2 excluding "aware" child participants.

Choice of particle

Figure C11a shows the proportion of speaker-appropriate (i.e., majority) particles produced by "unaware" adults in experiment 2 (n = 18) as a function of noun type and day. The model predicting production of the majority particle showed no significant intercept (b = -0.03, SE = 0.09, z = -0.30, p = .761), suggesting that performance was not greater than chance (50%), no effect of day (b = 0.04, SE = 0.04, z = 0.81, p = .421), noun type (b = -0.12, SE = 0.10, z = -1.25, p = .213), and no day by noun type interaction (b = -0.11, SE = 0.07, z = -1.62, p = .105).

Figure C11b shows the proportion of speaker-appropriate particles produced by "unaware" children in experiment 2 (n = 27) as a function of noun type and day. The model predicting production of the majority particle showed no significant intercept (b = 0.10, SE = 0.09, z = 1.10, p = .271), suggesting that performance was not greater than chance (50%), no effect of day (b = 0.05, SE = 0.04, z = 1.16, p = .244), noun type (b = -0.04, SE = 0.08, z = -0. 57, p = .568), and no day by noun type interaction (b = 0.05, SE = 0.05, z = 0.92, p = .358).

These analyses suggest that the partial speaker-identity cues were not learnt by "unaware" adults or children.

2AFC data

Unaware" adults' and children's 2AFC task performance in Experiment 2 across old and novel nouns is shown in Figure C12. The model predicting majority particle choice for adults showed no significant intercept (b = 0.15, SE = 0.12, z = 1.27, p = .203), suggesting that performance was not greater than chance (50%), and no significant main effect of noun

type (b = 0.14, SE = 0.26, z = 0.53, p = .594). The model predicting majority particle choice for children showed a significant intercept (b = 0.21, SE = 0.10, z = 2.21, p = .027), suggesting that performance was greater than chance (50%), and no significant main effect of noun type (b = -0.02, SE = 0.19, z = -0.10, p = .923). Thus, unlike adults, “unaware” children made speaker appropriate particle choices more frequently than expected by chance.

Figure Captions

Figure 1. Summary of tasks completed in each of the four experimental sessions.

Figure 2. Scene from a sentence training trial. Participants heard Katie produce a sentence describing the picture in her language (e.g., glim dog dak) and were asked to repeat it.

Figure 3. Correct particle usage by noun type (old, novel) and day (day 1, day 4) in the production task (Experiment 1). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%).

Figure 4. Correct particle choice by noun type (old, novel) in the two forced choice task (Experiment 1). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%).

Figure 5. Regularization performance in Experiments 1, 2 and 3. The first row shows entropy particle usage (with high entropy indicating more variable particle usage) for the two testing sessions, for productions involving old nouns, novel nouns, or taken across all nouns in the language. The dashed line indicates the expected entropy if participants were producing the two particles according to the trained frequencies. The second row shows lexical mutual information of particle usage, indicating the extent to which particle choice is conditioned on the noun, with high mutual information indicating more conditioned (less unpredictable) particle usage. The dashed line indicates the expected mutual information if participants were producing the two particles at random according to the trained frequencies. Error bars indicate bootstrapped 95% confidence intervals.

Figure 6. Proportion of majority particle usage by noun type (old, novel) and day (day 1, day 4) in the production task (Experiment 2). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%) and performance that matches the input (75%).

Figure 7. Proportion of majority particle choice by noun type (old, novel) in the two forced choice task (Experiment 2). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%) and performance that matches the input (75%).

Figure 8. Speaker identity mutual information of particle usage in Experiments 1, 2, and 3, indicating the extent to which particle choice is conditioned on the speaker, with high mutual information indicating more conditioned particle usage. Dashed lines indicate the expected mutual information if participants were producing the two particles entirely unconditioned on speaker identity; the solid line shows mutual information of the input set. Error bars indicate bootstrapped 95% confidence intervals.

Figure B9. Correct particle usage by noun type (old, novel) and day (day 1, day 4) in the production task (Experiment 1; unaware participants). Error bars indicate bootstrapped 95% confidence intervals.

Figure B10. Correct particle choice by noun type (old, novel) in the two forced choice task (Experiment 1; unaware participants). Error bars indicate bootstrapped 95% intervals.

Figure C11. Proportion of majority particle usage by noun type (old, novel) and day (day 1; day 4) in the production task (Experiment 2; unaware participants). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%) and performance that matches the input (75%).

Figure C12. Proportion of majority particle choice by noun type (old, novel) in the two forced choice task (Experiment 2; unaware participants). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%) and performance that matches the input (75%).

Table 1: *Mean age and Automated Working Memory Assessment (AWMA) task performance in Experiments 1 -3.*

		m; f	mono; biling	Age (yrs)	WordRec[a]	MazMem[a]	BackDigRec[a]
Exp. 1	Children	20; 10	15; 15	5.99 (0.35)	100.93 (14.76)	115.00 (15.10)	107.57 (20.78)
	Adults	8; 22	15; 15	21.26 (2.61)	91.29 (20.10)	100.00 (13.68)	105.77 (14.58)
Exp. 2	Children	13; 17	15; 15	6.12 (0.33)	102.90 (19.17)	110.20 (16.48)	103.30 (15.10)
	Adults	7; 23	15; 15	22.07 (3.55)	90.86 (17.22)	104.27 (12.82)	105.30 (16.32)
Exp. 3	Children	15; 15	15; 15	6.03 (0.45)	104.83 (14.89)	109.89 (17.23)	113.27 (18.11)
	Adults	13; 17	15; 15	22.88 (4.81)	92.33 (20.31)	104.27 (21.29)	107.30 (14.25)

Notes.

Standard deviations in parentheses.

[a]standard scores.

m= male; f = female; mono = monolingual; biling = bilingual; WordRec = Word Recall; MazMem = Mazes Memory; BackDigRec = Backwards Digit Recall

Table 2.

Performance in Experiment 1 by aware and unaware participants.

		% correct (*SD*s)	
	N	Production	2AFC
Adults			
Aware	30	.97 (0.16)	.99 (0.10)
Unaware	na	Na	na
Children			
Aware	12	.84 (0.37)	.83 (0.38)
Unaware	18	.84 (0.37)	.80 (0.40)

Notes.

2afc = 2 alternative forced choice. na = non-applicable

Table 3

Performance in Experiment 2 by aware and unaware participants.

		% correct (*SD*s)	
	n	Production	2AFC
Adults			
Aware	12	.71 (0.45)	.77 (0.42)
Unaware	18	.50 (0.50)	.54 (0.50)
Children			
Aware	3	.60 (0.49)	.60 (0.49)
Unaware	27	.52 (0.50)	.55 (0.50)

Notes.

2AFC = 2 alternative forced choice.

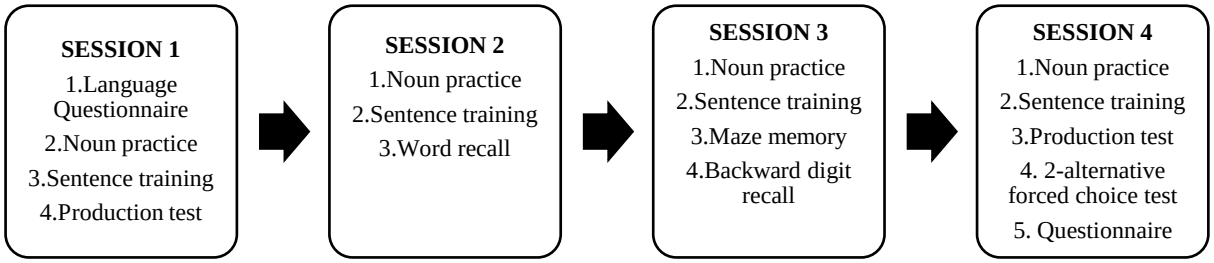

Figure 1. Summary of tasks completed in each of the four experimental sessions.

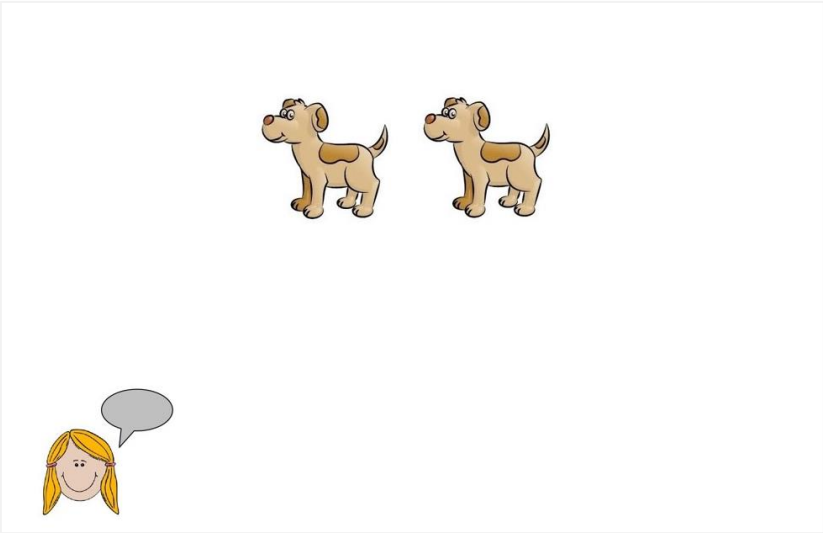

Figure 2. Scene from a sentence training trial. Participants heard Katie produce a sentence describing the picture in her language (e.g., glim dog dak) and were asked to repeat it.

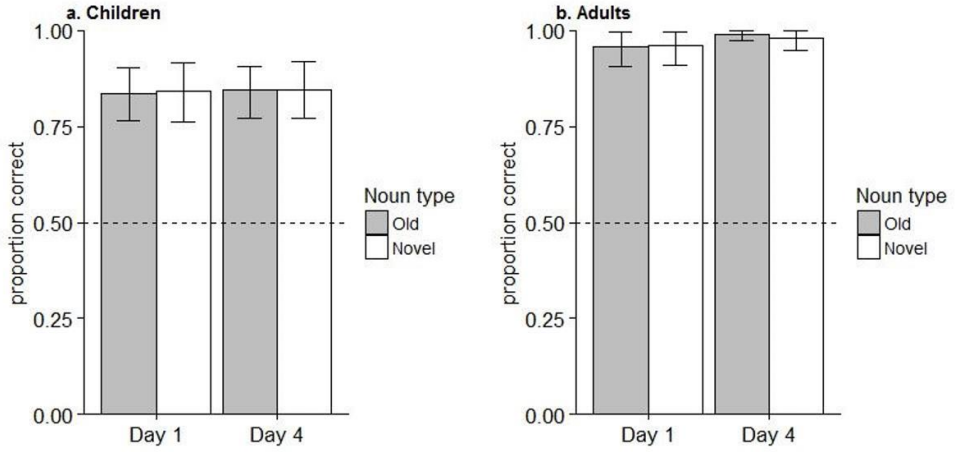

Figure 3. Correct particle usage by noun type (old, novel) and day (day 1, day 4) in the production task (Experiment 1). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%).

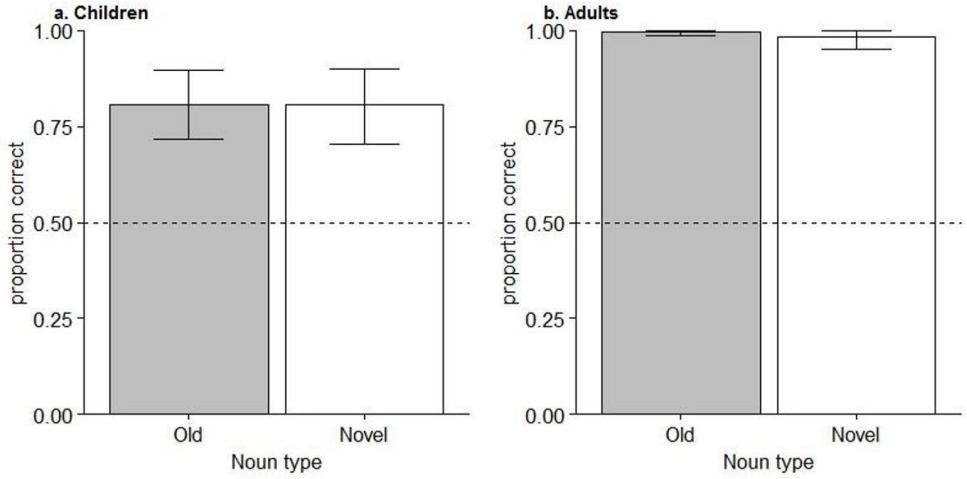

Figure 4. Correct particle choice by noun type (old, novel) in the two forced choice task (Experiment 1). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%).

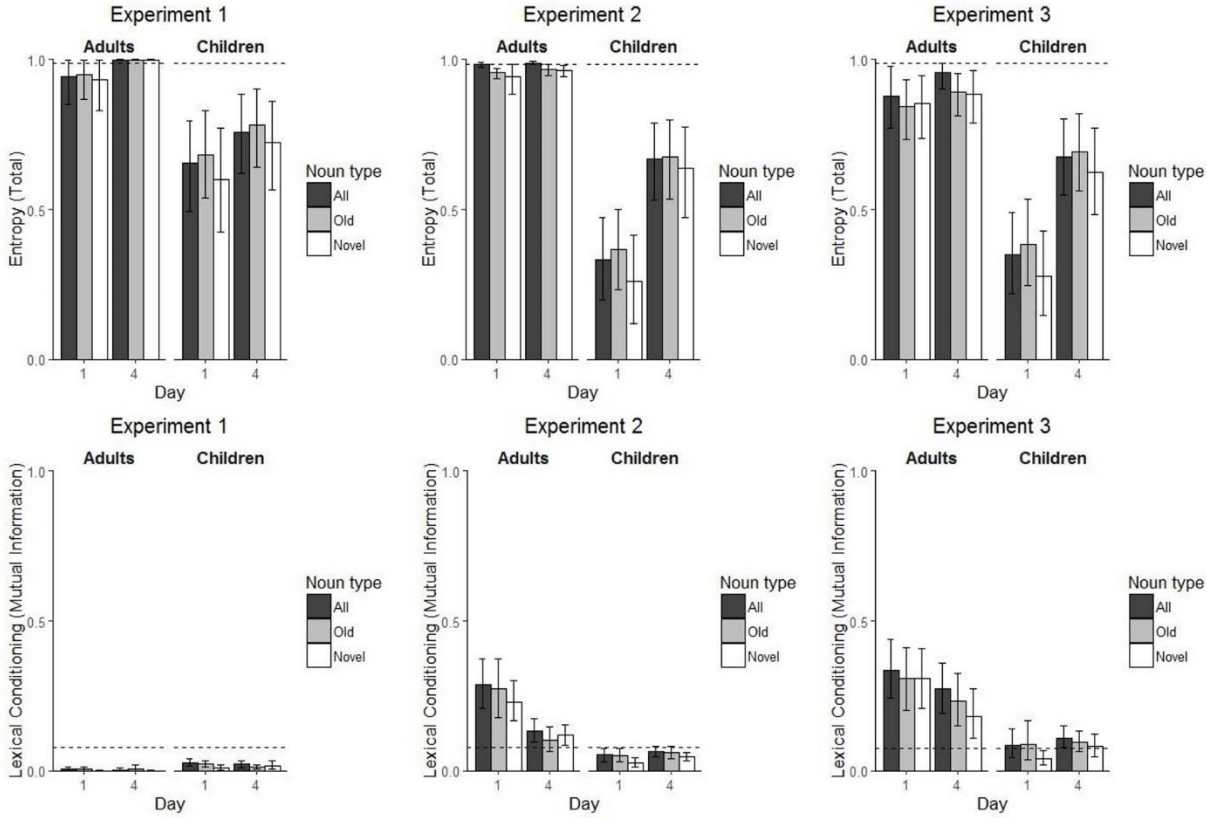

Figure 5. Regularization performance in Experiments 1, 2 and 3. The first row shows entropy particle usage (with high entropy indicating more variable particle usage) for the two testing sessions, for productions involving old nouns, novel nouns, or taken across all nouns in the language. The dashed line indicates the expected entropy if participants were producing the two particles according to the trained frequencies. The second row shows lexical mutual information of particle usage, indicating the extent to which particle choice is conditioned on the noun, with high mutual information indicating more conditioned (less unpredictable) particle usage. The dashed line indicates the expected mutual information if participants were producing the two particles at random according to the trained frequencies. Error bars indicate bootstrapped 95% confidence intervals.

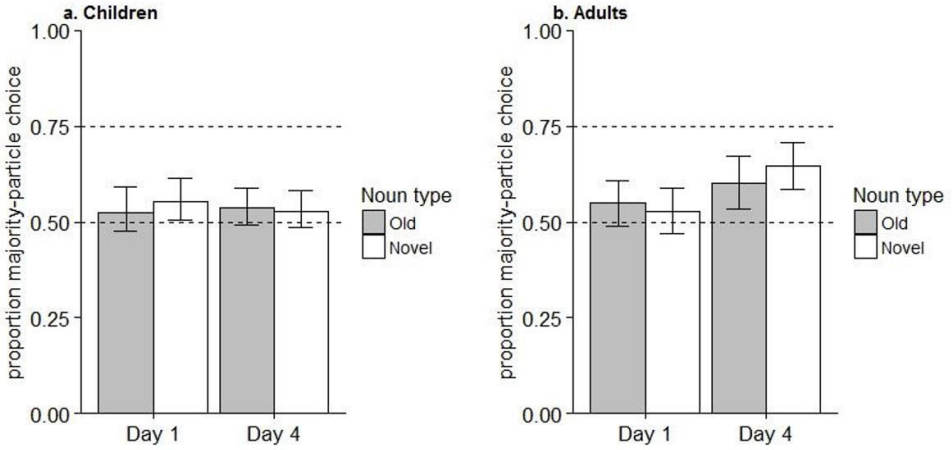

Figure 6. Proportion of majority particle usage by noun type (old, novel) and day (day 1, day 4) in the production task (Experiment 2). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%) and performance that matches the input (75%).

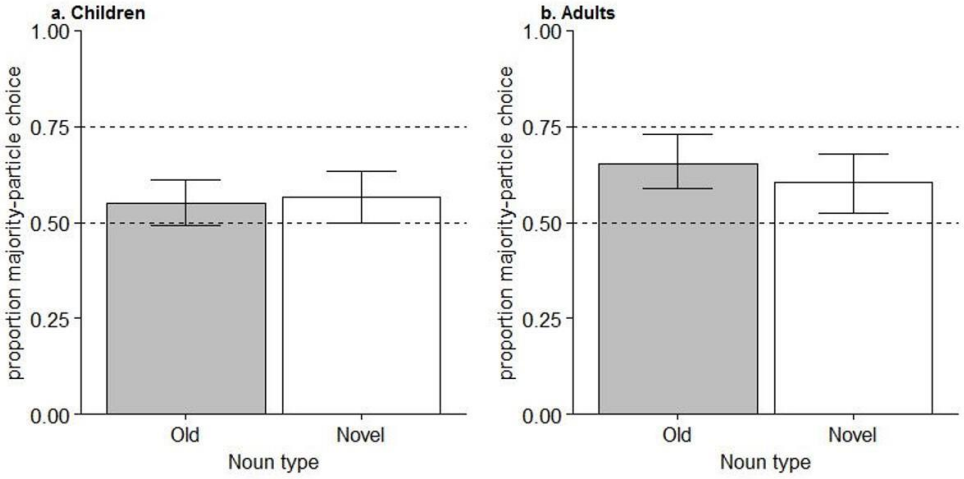

Figure 7. Proportion of majority particle choice by noun type (old, novel) in the two forced choice task (Experiment 2). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%) and performance that matches the input (75%).

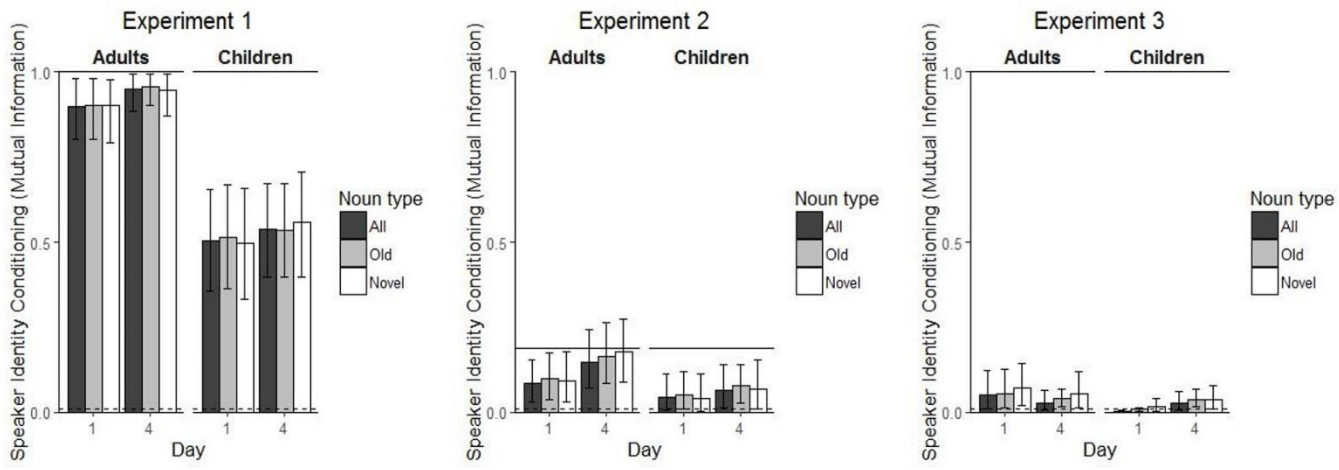

Figure 8. Speaker identity mutual information of particle usage in Experiments 1, 2, and 3, indicating the extent to which particle choice is conditioned on the speaker, with high mutual information indicating more conditioned particle usage. Dashed lines indicate the expected mutual information if participants were producing the two particles entirely unconditioned on speaker identity; the solid line shows mutual information of the input set. Error bars indicate bootstrapped 95% confidence intervals.

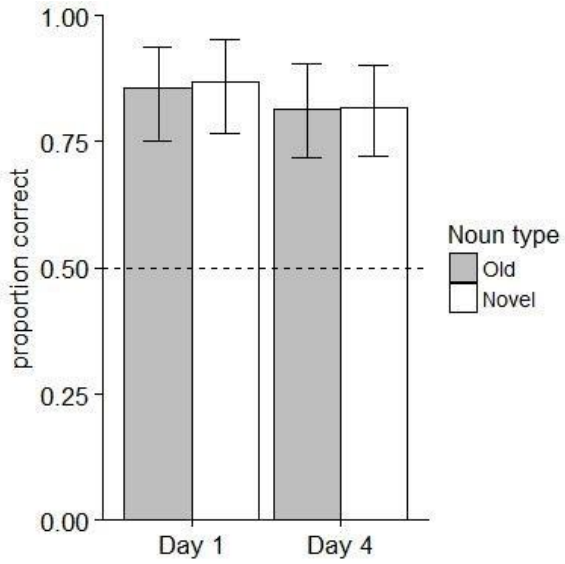

Figure B9. Correct particle usage by noun type (old, novel) and day (day 1, day 4) in the production task (Experiment 1; unaware participants). Error bars indicate bootstrapped 95% confidence intervals.

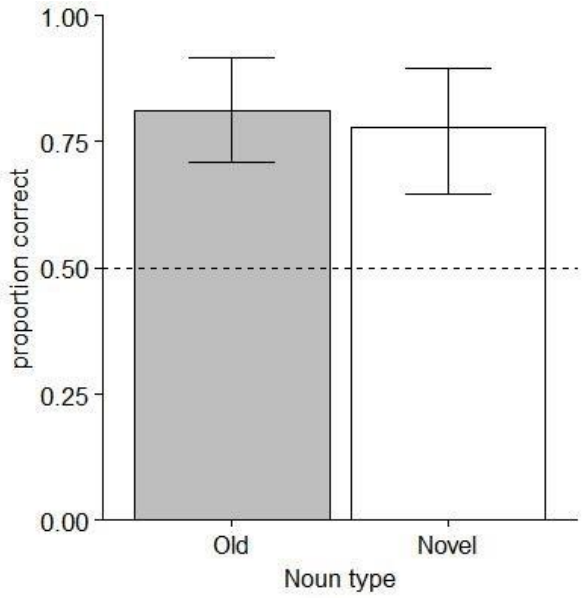

Figure B10. Correct particle choice by noun type (old, novel) in the two forced choice task (Experiment 1; unaware participants). Error bars indicate bootstrapped 95% intervals.

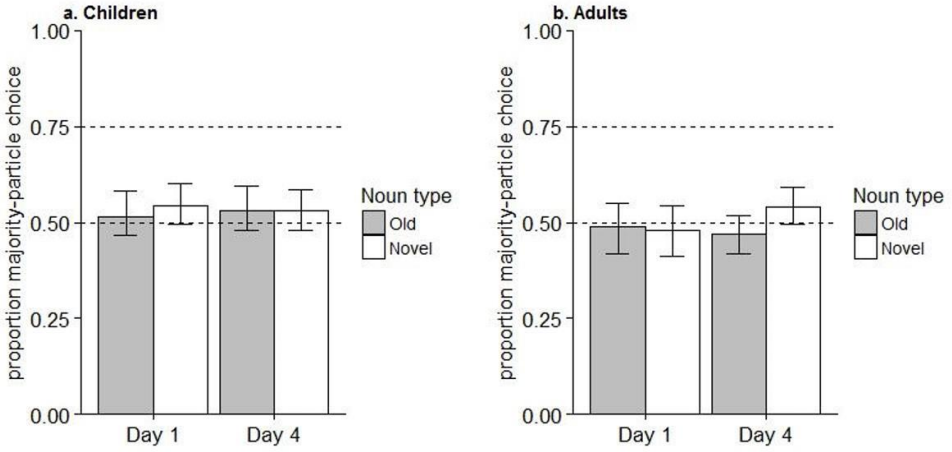

Figure C11. Proportion of majority particle usage by noun type (old, novel) and day (day 1; day 4) in the production task (Experiment 2; unaware participants). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%) and performance that matches the input (75%).

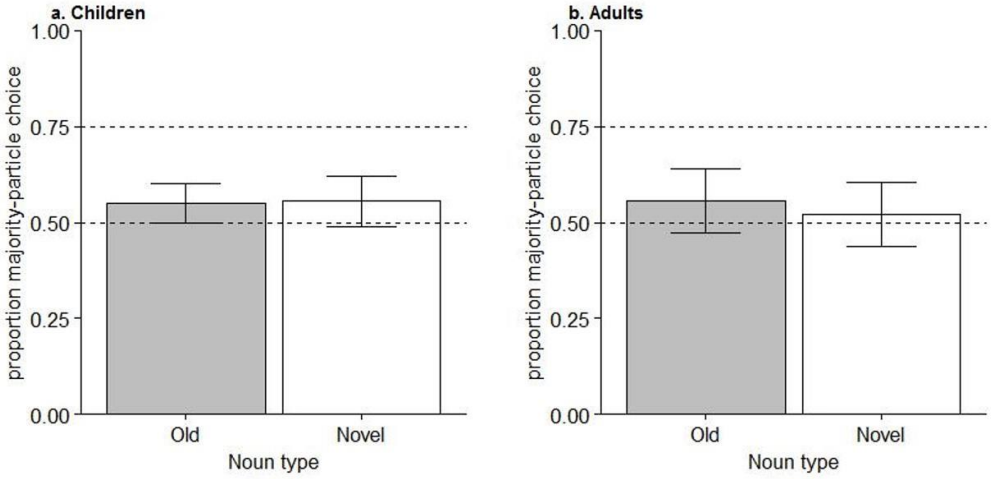

Figure C12. Proportion of majority particle choice by noun type (old, novel) in the two forced choice task (Experiment 2; unaware participants). Error bars indicate bootstrapped 95% confidence intervals. Dashed lines show chance performance (50%) and performance that matches the input (75%).

Highlights

- We tested children's and adults' ability to acquire socially conditioned variation.
- Variable particle usage in a semi-artificial language was cued by speaker identity.
- Both children and adults picked up on the sociolinguistic speaker identity cue.
- Any tendency to regularize did not prevent the learning of conditioned variation.