Testing the Tolerance Principle:
Children form productive rules when it is
more computationally efficient

Kathryn D. Schuler[a,b], Charles Yang[a] and Elissa L. Newport[b]

[a]University of Pennsylvania

[b]Center for Brain Plasticity and Recovery, Georgetown University

Acknowledgements

This research was supported in part by NIH Grant R01 HD037082, by the Feldstein Veron Innovation Fund, and by the Center for Brain Plasticity and Recovery at Georgetown University.

Data and code available at: https://github.com/pennchildlanglab/tolerance-principle-project

Abstract

During language acquisition, children must learn when to generalize a pattern – applying it broadly and to new words ('add –ed' in English) – and when to restrict generalization, storing the pattern only with specific lexical items. But what governs when children will form productive rules during language acquisition? How do they determine when a pattern is widespread enough to generalize to novel words, and when a pattern should not extend beyond the cases they have observed in their input? One effort to quantify the conditions for generalization, the Tolerance Principle (Yang, 2016), has been shown to accurately predict children's generalization behavior in dozens of corpus-based studies. The Tolerance Principle hypothesizes that a general rule will be formed when it is computationally more efficient than storing lexical forms individually. Here we test the Tolerance Principle in two artificial language experiments with children. In both experiments, we exposed children to a language with 9 novel nouns, some of which followed a regular pattern to form the plural (-ka) and some of which were exceptions to this rule. As predicted by the Tolerance Principle, in Experiment 1 we found that children exposed to 5 regular forms and 4 exceptions generalized, applying the regular form to 100% of novel test words. Children exposed to 3 regular forms and 6 exceptions did not extend the rule, even though the regular form was still the majority token in this condition. In Experiment 2, we found that children continued to behave categorically: either forming a productive rule (applying the regular form on all test trials) or using the regular form no more than predicted by chance. We found that the Tolerance Principle can be used to predict whether children will form a productive generalization or not based on each child's individual vocabulary size. The Tolerance Principle appears to capture something fundamental about the way in which children form productive generalizations during language acquisition.

1. Introduction

When children learn a language, they do not simply memorize the words and sentences they are exposed to; they acquire the patterns by which words and sentences are formed. For example, in English, we add -*ed* to form the past tense of verbs (e.g. *walked, jumped, played*) and -*s* to make nouns plural (e.g. *apples, dogs, friends*). While there are exceptions to these processes — the past tense of *go* is *went*, the plural of *goose* is *geese* — the overwhelming majority of words obey these regular, productive patterns.

We call these patterns *productive* because they generalize broadly to new or novel words. When new verbs enter our vocabulary, we apply these productive patterns to generate their inflected forms (e.g. *googled* and *instagrammed*). When we make speech errors, it is often because we have mistakenly applied a productive pattern to a word that is actually a lexical exception (e.g. accidentally saying "breaked" as the past-tense of "broke") (Stemberger, 1983). In child language, evidence for the productivity of these patterns is even more striking. As children acquire their native language, they often overgeneralize the regular patterns, applying them in ways their language models do not: "Daddy goed to the store" or "I have two foots!" (Marcus et al., 1992; Maslen et al., 2004; Pinker, 1995; Yang, 2002). Further, when asked to inflect completely novel words in experiments — for example, "This is a wug. Now there are two of them, there are two ____." — children overwhelmingly apply the regular form (Berko, 1958).

While evidence shows that children acquire these regular, productive patterns, not all patterns in language are productive. Some inflected forms are idiosyncratic to a single lexical item — the exceptions like *go-went* — but languages also contain patterns that apply only to a restricted subset of items (e.g. *sing/sang/sung* and *ring/rang/rung*). These irregular patterns are not productive, in that they do not generalize broadly to novel words.

Though this aspect of productivity is less often discussed, there is substantial evidence to support the claim that irregular forms are not productive. For example, in rating and production task experiments, adults generalize by analogy to an irregular form only under very limited circumstances (Albright & Hayes, 2003; Ambridge, 2010; Bybee & Moder, 1983; Bybee & Slobin, 1982; Prasada & Pinker, 1993), and children almost never do. Children prompted to provide the inflected form of *gling*, for example, nearly always produce *glinged*, not *glang*, even though *gling* is analogous to sing and ring (Berko, 1958). Further, while it is common for children to overgeneralize regular forms during the acquisition process (Marcus et al., 1992; Maslen et al., 2004; Pinker, 1995) — some estimate nearly 10% of children's productions contain such an overregularization (Hoeffner, 1997; Maratsos, 2000; Maslen et al., 2004; Yang, 2002) — children rarely overgeneralize irregular forms. Xu & Pinker (1995), for example, found overgeneralization of irregular forms in only 0.02% of children's productions. A similar tendency to overgeneralize regular but not irregular patterns has also been observed in German (Clahsen & Penke, 1992), Inuktitut (Allen, 1996), Spanish (Caprin & Guasti, 2009), Bantu languages (Demuth & Nurse, 2003), Swahili (Deen, 2005), and many others (see Yang, 2016, for a review).

The tendency to generalize some rules but to restrict others motivates the question: what governs when children will form productive rules during language acquisition? How do they determine when a pattern is widespread enough to generalize to novel words, and when a pattern should not extend beyond the cases they have observed in their input?

1.1 Previous Approaches to Rule Learning

Rule learning in language acquisition has been investigated for many years, but most work has focused on the evidence for abstract rules; relatively little research has investigated the question of productivity of regular versus irregular rules.

Some have argued that regular and irregular patterns can be acquired without abstract rules. Rumelhart & McClelland (1986) proposed one such no-rules approach, arguing that children could learn to inflect verbs correctly via a simple pattern association mechanism. To demonstrate, they trained a connectionist model to associate features of a verb stem with features of the verb's past tense form. To elicit overgeneralizations at an appropriate developmental stage, Rumelhart & McClelland's model increased the number of regular forms the model received at this time. While this and similar models have been shown to be somewhat successful at producing regular English past tense forms in response to their present tense forms, the overall pattern of acquisition does not match what we observe in children. Connectionist models overgeneralize regular forms at approximately the same rate as children do, but they overgeneralize irregular forms much more often (Cazden, 1968; Kuczaj, 1977; Marcus et al., 1992).

In contrast to the nothing-is-rules approach, some approaches have argued that all inflected forms are generated by rules and that, in many circumstances, more than one rule can apply (e.g. (Chomsky &

Halle, 1968; M. Halle & Marantz, 1993; Kiparsky, 1982; Mohanan, 1986; Yang, 2002). These rules will differ, however, in how many lexical items or what type of lexical class they apply to. Thus, for example, to form the past tense of *go*, both goed and went will be generated. Competition between rules is resolved by stating that a more specific rule will be preferred over a more general one (Anderson, 1969; Brown & Hippisley, 2012; M. Halle & Marantz, 1993; Stump, 2001) or by organizing the lexicon in levels (also called *strata*) and stating that rules applied on earlier levels take precedence over those applied later (Halle & Mohanan, 1985; Kiparsky, 1982; Mohanan, 1986). Under such accounts, irregular rules apply to irregular lexical items like *ring* because they are more readily available for those lexical items, either because they are more specific or because they apply on an earlier level in the derivation.

Others have argued that some forms can be generated by rules, but others must be memorized. For example, *dual-route* approaches — in which the regular forms are generated by abstract rules while the irregulars and exceptions are stored in memory (Clahsen 1999; Pinker 1999) — argue that the regular form can only apply if there is no irregular form to block it. When both regular and irregular forms are available, the competition between them is typically resolved via some variation of the Blocking-and-Retrieval Failure hypothesis (Marcus et al. 1992): the irregular form applies whenever available, *blocking* the regular rule. If the learner fails to retrieve an irregular (*retrieval failure),* then nothing blocks the regular form and it is free to apply. Proponents of this account point out that children's errors are well-predicted by this approach. If no irregular form has been acquired, children are predicted to make an overgeneralization error, applying the regular form. If neither an irregular nor the regular form have been acquired, then children are predicted to make an error of omission, producing a bare stem with no inflection.

Still others argue that the so-called regular and irregular forms are both *constructions*, not abstract rules. Under accounts of *entrenchment*, the more often a learner hears a construction, the more entrenched that construction becomes, and the learner becomes less and less likely to use another form (Ambridge et al. 2008; Stefanowitsch 2008). An extension of the entrenchment proposal involves *statistical preemption*, in which children make predictions about what constructions will occur in a given context and take their non-occurrence — more specifically, the occurrence of something else — as negative evidence that their predicted forms were ungrammatical (Goldberg 1995; Robenalt and Goldberg 2015). This approach is reminiscent of the concept of *blocking* described above: children learn to use "ran" instead of "runned" because every time they predicted the “runned” construction, they heard “ran” instead. Goldberg (2016) has argued that exemplars sharing surface forms or functions cluster together into dynamic “constructional categories”, and a construction is generalized (i.e., extended to novel contexts) when the category is sufficiently well attested (although this is not precisely defined). Statistical preemption is also invoked here, in the sense that “learners learn to avoid using one construction, even when the construction’s constraints would seem to be satisfied, if an alternative formulation has been sufficiently witnessed instead” (Goldberg 2016).

Interestingly, each of these (otherwise very different) accounts rely heavily on the same idea: that the more readily available form ‘wins’. While each offers a reasonable framework with which to think about regulars, irregulars, and the competition between them, none details a mechanism that governs how or when a process becomes productive. Most approaches have focused on predicting whether children will apply the regular rule or an irregular form for any given stem, but they have not addressed precisely how children acquire the regular (productive) form from the linguistic input they receive or determine whether a particular form is indeed the regular productive form (or in the case of constructions, what counts as sufficiently well attested).

1.2 Previous approaches to productivity

While the rule learning literature has not focused directly on the question of productivity, many other linguists and language acquisition researchers have proposed metrics for determining the degree to which a linguistic pattern will be productive. Some approaches use simple computations based on the type or token frequencies of the competing morphological forms. For example, Bybee (1995) argued that productivity could be predicted by the number of lexical items on which the affix can occur. Aronoff (1976) proposed a slightly more complex metric, later formalized by Baayen & Lieber (1991), in which the productivity of a given form (the Word-Formation Rule) could be quantified as the number of lexical items which take that form, divided by the number of lexical items that could potentially take that form. More complex models attempt to optimize the processing cost of using various forms or inferring the regular form from aspects of its distribution. Taatgen & Anderson (2002), for example, proposed a model based on the

ACT-R framework (Anderson and Lebiere 1998) in which several inflection strategies compete to produce the English past tense and the model selects whichever requires the least effort. Another productivity-as-optimization model, O'Donnell's Fragment Grammars, (2011; 2015), uses probabilistic inference to generate a hypothesis about which is the optimal account of the observed data: reusable stored forms, a productive generalization, or some combination of both.

While these approaches are on the surface quite different, nearly all share important assumptions about the nature of productivity — assumptions that often go unquestioned in the literature. For example, nearly all are based on the assumption that statistical dominance forms the basis for productivity, an idea that has been emphasized in the productivity literature for decades. For instance, a classic text states that "(a) form which is statistically predominant is also likely to be productive for new combinations" (Nida, 1949, 45), and the regular suffix -d in English is explicitly identified with statistical majority (Bloch, 1947). The approaches reviewed above clearly differ in precisely how they identify the majority form, but all invoke the idea of statistical dominance in one way or another. Some explicitly state that the more types or tokens a form applies to, the higher its "productivity index" and the more likely it is to be productive (Bybee, 1995; Aronoff, 1976; Baayen & Lieber 1991). Others build the notion of statistical dominance into the parameters of more complex probabilistic models (Taatgen & Anderson, 2002; O'Donnell 2011, 2015).

Proposals in the psychological study of language acquisition similarly emphasize the role of statistical dominance. Early research on children's morphological acquisition suggests that children initially store specific exemplars until enough have been accumulated to induce a rule (Ivimey, 1975; Kuczaj, 1977; MacWhinney, 1978). Interestingly, the connectionist approach to generalization is based on this same idea For Rumelhart and McClelland (1986), the model was initially presented with a few epochs of training data: each had 10 verbs, of which 8 were irregular. After the model had stored these items reliably, a much larger sample of 420 verbs was presented, of which a dominant majority were regular (334). It was at this point that the network learned the productivity of "-ed" and began over-regularizing irregular verbs. Thus productivity in this model is again the result of statistical dominance, which has become known as the Critical Mass Hypothesis (Marchman and Bates, 1994): a critical mass of regular verbs is necessary for the generalization of "-ed".

The research on productivity, especially in language acquisition, has been dominated by the case of English past tense, where the majority form does correctly identify the productive "-ed". However, across a broader range of cases in English and other languages of the world, mere statistical dominance does not successfully identify which forms are productive and which are not. In some cases, even a very dominant majority form fails to become the general rule. For example, over 80% of English words stress the first syllable (Cutler & Davis, 1987), yet the lexical stress system of English does not treat initial stress as a general rule with the remaining words' stress patterns memorized as exceptions. Instead, stress assignment is conditioned on syntactic categories (e.g., nouns and verbs; Chomsky & Halle, 1968; Kelly & Bock 1988) as well as syllable structure (Guion et al. 2003). In other cases, a rule that covers very few words can nevertheless be productive. For example, the German noun plural system has five suffixes; of these, -s is attached to only about 4% of the nouns, yet it productively applies to certain types of nouns (Clahsen 1999).

Beyond statistical dominance, the productivity metrics we have reviewed so far also rely on the assumption that some form will emerge as productive. While this winner-take-all approach is a popular strategy in the productivity literature, it is important to point out that some aspects of language have no productive rules. Such phenomena are known as "gaps": inflected forms of some words are simply unavailable (Baerman et al. 2010). For example, there are about seventy verbs in Russian that lack an acceptable first-person singular nonpast form (Halle 1973; Sims 2006): *lažu 'I climb', * *deržu 'I talk rudely', etc. are rejected by native speakers even though there is nothing semantically or phonologically deviant about them. Even morphologically simple languages such as English contain such gaps: the past participle of stride is unavailable for speakers of most English dialects who reject all three potential forms (*I have strode/stridden/*strided; Pullum & Wilson 1977). In the study of language acquisition, it has been shown that in Polish, masculine nouns in the singular genitive either take an -*a* or -u suffix, which appear with approximately 65% and 35% of the nouns. But neither is productive (Dabrowska 2001), and learners must resort to lexically specific memorization for each noun. In other words, in the case of gaps, a learning model must *fail* to generalize in order to provide a successful account of child language acquisition.

Finally, while researchers may be capable of using the metrics reviewed here to estimate the productivity of a morphological process in natural language, it is unlikely that young children have access to the complex variables that underlie these computations. And even if one assumes that they do, it is not clear how a child would apply the resulting estimate of productivity to decide whether or not a rule should be productive. The computations that underlie each of these quantitative approaches typically result in an index of productivity (e.g., a value between 0 and 1), which assumes that rules or patterns can have some degree of productivity. What is the threshold index for which children can assume a rule applies to novel situations? An index of 0.50? An index of 0.75?

To summarize, many different approaches to rule learning and productivity share the underlying assumption that the more accessible form wins — and we agree with this general notion. However, detailed formal accounts of precisely when a child will form a productive generalization have so far relied on several problematic assumptions about the nature of productivity in language and its acquisition. When one looks carefully at these assumptions, it is clear that what is missing from the literature is an account of children's generalization behavior that can (1) successfully identify which forms are productive in a range of languages (beyond the English past tense), (2) account for the many circumstances in which no productive form exists, and (3) provide an evaluation metric with which children can determine, based on the language input they have received, whether or not a rule is productive. In the present paper, we focus on one such model that fulfills these requirements: the Tolerance Principle (Yang, 2016).

1.3 The Tolerance Principle

The Tolerance Principle (Yang, 2016, 2005) is based on a learning model that quantifies the precise conditions for generalization or productivity during language acquisition. It hypothesizes that a general productive rule will be formed when doing so is computationally more efficient than storing lexical forms individually. The model determines which option is more efficient by calculating the time complexity required for applying a rule, compared with the time complexity required for accessing individual lexical forms. To illustrate, imagine that a learner is faced with a potential rule – for example, the English 'add –ed to make a verb in the past tense.' The English learner has encountered many items that obey this rule (regular forms) as well as many items that do not (irregular forms or exceptions). To be maximally efficient in forming the past tense of verbs in her language, the learner can do one of two things:

(1) **Store all lexical forms individually**: store every past tense item individually in a list ranked by frequency, searching the list every time there is an occasion to express the past tense of a verb.

(2) **Form a productive rule**: store only the exceptions in a frequency-ranked list. To express the past tense, the learner searches the list of exceptions first. If the target verb is not among these exceptions, the learner applies the rule 'add –ed.'

Formally, if *R* is a rule that may apply to *N* lexical items and there are *e* exceptions to this rule, the time required to access the rule can be expressed as *T(N, e)*. If *R* is productive (as in (2) above) then the rule is not applied until the learner has first evaluated and rejected every exception (*e*) on the list. In other words, applying a productive rule consumes *e* units of time. The time required for exceptions, on the other hand, is determined by the frequency of the lexical item (i.e., its rank in the list of exceptions). To compute the time complexity *T(N, e)*, Yang (2016, p. 48) calculates "the weighted average of time units over the probabilities of these two sets of items." If *R* is unproductive (as in (1) above) then all *N* items are treated as exceptions and are listed in order of frequency. The time complexity under these circumstances is expressed *T(N, N)*, as the number of exceptions *e* is equivalent to the number of items in the list *N*. The learner compares the time complexity required to form a productive rule, *T(N, e)*, with the time complexity required when all *N* items are stored as lexical exceptions, *T(N, N)*. By solving this equation for *e*, the Tolerance Principle computes the precise number of exceptions that a productive rule can tolerate before forming a rule becomes the less efficient strategy. This solution is as follows:

(3) **Tolerance Principle:** Let R be a rule that is applicable to N items, of which e are exceptions. R is productive if and only if: $e \leq \theta N = N/\ln(N)$

In other words, it is only more efficient to form a productive rule when the number of exceptions is less than the number of items divided by the natural log of the number of items. To illustrate, consider a category of 9 items. Given a rule *R* that may apply to these 9 items, the Tolerance Principle predicts that 4.096 (9/ln9) exceptions will be tolerated before forming a productive rule becomes less efficient than storing individual items. This means that learners will form a productive rule if there are 4 or fewer exceptions to the rule, but not if there are 5 or more. Importantly, this implies that the distinction between forming a productive rule and storing individual lexical items is a categorical one. There is a theoretical tipping point at which forming a productive rule becomes less computationally efficient than the alternative strategy. The Tolerance Principle allows us to compute this tipping point.

Yang's Tolerance Principle has been shown to predict generalizations in corpus data from dozens of regular and irregular patterns in a wide range of languages (see, e.g., Fernández-Dobao and Herschensohn 2019; Merkuur et al. 2020; Garcia 2019; Labov 2020; Bjornsdottir, 2021 for recent application to a typologically diverse range of languages). For example, to test the Tolerance Principle on the English past-tense, Yang (2016) analyzed 5 million words of child-directed English from CHILDES (MacWhinney, 2000) and found 1022 unique past-tense verbs. In a class of 1022, the Tolerance Principle predicts that the '*add -ed*' rule should tolerate 147 exceptions, and indeed Yang found only 127. The irregular patterns in English, however, do not fare so well. Even the irregular class that has the highest homogeneity — the *ing-ang* class (as in sing-sang and ring-rang) — has too many exceptions. The CHILDES English input corpus has 8 verbs that end in *-ing*, but only three change the past tense to -ang (*ring*, *sing*, *spring*) — the remaining five do not (*bring, fling, sting, swing, wing*). Since 5 exceeds the threshold of productivity for a class of 8 verbs {theta}, the *ing-ang* pattern is predicted to be unproductive, just as it appears to be in children's speech.

1.4 The Present Experiments

The Tolerance Principle accurately predicts children's generalizations in corpus data across many languages, demonstrating that it is a viable model of productivity in language acquisition. However, in order to have adequate data for testing predictions, corpus analyses combine data from multiple children at different ages, not all of whom may show the same behaviors; and one can only test those patterns that happen to occur in the sample. In the present experiments, we apply two well-known acquisition paradigms — artificial language learning and the "wug" test — to submit the Tolerance Principle to further experimental scrutiny. The artificial language paradigm allows us to control the input to child learners, manipulating the precise number of lexical items that obey a rule or are exceptions. Across two experiments, we exposed children to either a language in which the Tolerance Principle predicts that learners should form a productive rule (as in the –ed example above), or one in which it predicts that learners will not form a productive rule (as in the *ing*-*ang* example above). We ask whether the Tolerance Principle correctly predicts when a pattern in an artificial language is widespread enough for a child to form a productive rule, using a "wug" test to assess whether children have formed a productive rule (one that applies to novel lexical items) or have restricted generalization. In Experiment 1, we exposed children to languages in which the regular form applied to the most frequent nouns in the input — allowing us to ask whether children form a productive generalization according to the number of types that take the regular form, as predicted by the Tolerance Principle, or if they simply regularize the most frequent token in the input. In Experiment 2, we asked whether the Tolerance Principle can predict children's generalization behavior, even when the regular form is not applied to the most frequent nouns in exposure. In both experiments, we test two further predictions. First, the Tolerance Principle, in contrast to many other metrics of productivity, predicts a categorical outcome: children will either form a productive generalization or they will not. We make this same precise prediction for children in our experiments. Second, since children's generalization behavior has been shown to differ substantially from adults in other contexts (e.g. Hudson Kam & Newport, 2005), we included adult controls to test this hypothesis for the Tolerance Principle as well.

2. Experiment 1

Our primary goal is to determine whether the Tolerance Principle accurately predicts when children acquiring an artificial language will form a productive rule. To that end, in Experiment 1 we aim to determine whether children form productive rules based on the number of *types* that take the regular form — as the Tolerance Principle predicts — or whether they do so simply based on the most frequent *token* in the input. While the idea of productivity based on type, not token frequency is not necessarily disputed — most theories of regular and irregular morphology accept the elevated importance of type over token frequency in productivity (e.g. Baayen & Renouf, 1996; Bybee, 1995; Pinker, 1991; Plunkett & Marchman, 1991; Rumelhart & McClelland, 1986) — differentiating between type and token frequency is especially important in our artificial language experiment, given that similar experiments find learners behave in systematic ways based on *token* frequencies (Austin, Schuler, et al, under review; Hudson Kam & Newport, 2005, 2009).

While there are important differences between these experiments and the present work – most notably that every type is probabilistically marked in those experiments, while here each type is marked with a consistent form – it is important to determine whether children will track type frequency in the present experiment as the Tolerance Principle suggests, or whether they will simply regularize the token that appears most often in the language input, as children in similar artificial language experiments have been shown to do.

In Experiment 1, we investigate this type-token distinction by pitting type and token frequencies against one another. We manipulate the number of types that take the regular form such that the Tolerance Principle predicts children should form a rule in one condition (in which there are 5 regulars and 4 exceptions), but not in the other (3 regulars and 6 exceptions). At the same time, we keep the token frequency of the regular form high in both conditions — at least 59% of tokens — such that children forming a rule based on the most frequent token should do so in both conditions.

We include adult participants in order to address whether the Tolerance Principle is a model of productive rule formation that applies only to children, or whether this model predicts adult behavior as well.

2.1 Method

Participants

Participants were 15 children (mean = 7.55 years, sd = 0.86 years, range = 6.05–8.89 years) and 20 adults. We excluded a further seven children from our analysis for failing to finish the experiment (3) or failing to produce the correct noun on at least 50% percent of production trials (4). Children were recruited from schools, camps, and day-cares in the Washington D.C. metro area and adults were recruited on Amazon Mechanical Turk.

All participants were native English speakers with normal to corrected-to-normal hearing and vision. Children received stickers and a set of small toys for their participation and their caregivers received a $10 travel reimbursement if they traveled to our lab. Adults were paid $10 for their participation.

Stimuli

To assess whether children follow the Tolerance Principle, we created an artificial language with one verb (*gentif*), 9 nonsense nouns (*mawg, tombur, glim, zup, spad, daygin, flairb, klidam, lepal*), and one rule (add *ka* to make nouns plural). In a category of 9 nouns, the Tolerance Principle predicts children should tolerate 9/ln(9) exceptions — that is, they should form a productive rule when there are 4.096 exceptions or fewer. To test this, we created one condition in which 5 nouns take the regular form and 4 are exceptions (below the 4.096 threshold) and one in which 3 nouns take the regular form and 6 are exceptions (above the 4.096 threshold). If children follow the Tolerance Principle, only those in the 5 regular, 4 exception condition should form a productive rule.

To create our exposure corpus, we assigned each noun a plural marker that either followed the rule (add *ka*) or was an exception (add *po, tay, lee bae, muy,* or *woo*), depending on the condition. Each noun appeared in an unmarked singular sentence (e.g., *gentif mawg*) and a marked plural sentence (e.g. *gentif mawg ka*). Each sentence was paired with a corresponding picture: singular sentences had one image of the corresponding object while plural sentences had 2, 4, or 6 (see Figure 1). All sentences began with our single verb *gentif*, meaning “there is/are”.

We generated 72 exposure sentences (24 singular, 48 plural) by varying noun frequency along a Zipfian distribution (Zipf, 1949), meaning the second most frequent noun was presented about half as often as the first, the third was presented about half as often as the second, and so on. The Zipfian distribution is important to our design for two reasons. First, the distribution of word frequency in natural language is approximately Zipfian, and the derivation underlying the Tolerance Principle assumes word frequencies follow this pattern. Second, by applying the regular form to the most frequent nouns in our Zipfian distribution, we created a situation in which type and token frequencies were in conflict. If children form productive rules based on the number of types that take the regular form, as the Tolerance Principle predicts, they should form a productive rule in the 5 regular, 4 exception condition, but not in the 3 regular, 6 exception condition. If children instead track the token frequency of the regular form in their input, they should form a productive rule in both conditions.

Table 1. Frequency of nouns and their plural markers in Experiment 1, Language A. The regular form appears in bold.

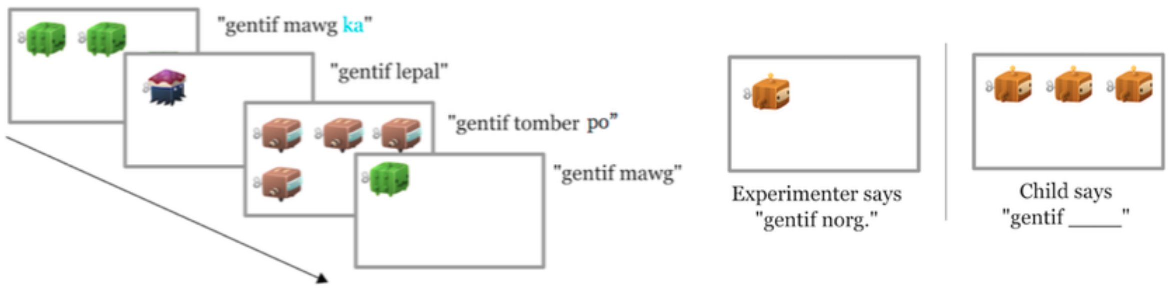
Rank	Freq	N Plural	Noun	5R4E	3R6E
1	24	16	mawg	**ka**	**ka**

2	12	8	tombur	**ka**	**ka**
3	8	5	glim	**ka**	**ka**
4	6	4	zup	**ka**	po
5	6	4	spad	**ka**	lee
6	4	3	daygin	po	bae
7	4	3	flairb	lee	tay
8	4	3	klidam	bae	muy
9	4	3	lepal	tay	woo

Finally, to ensure there was nothing idiosyncratic about the particular noun frequency ranking, we created two counterbalanced languages — A and B — that differed only in which nouns were paired with which markers (and thus, which nouns were more frequent).

Procedure

The experiment consisted of three parts: exposure to the language, a production test, and a rating test. During exposure, we presented participants with the 72 exposure sentences and their corresponding pictures in random order. On each trial, participants saw a picture - 1, 2, 4, or 6 instances of a noun - and were presented with the corresponding singular (for 1) or plural (for 2, 4, or 6) auditory sentence (see Figure 1a). Participants were instructed to repeat the sentence aloud (or type it into a response box on Mechanical Turk) before moving on to the next trial. We gave participants a short break every 18 trials, and children were offered a sticker during these breaks to encourage them to continue.



(a) Four example exposure trials

(b) Example production test trial

Figure 1. Examples of (a) four exposure phase trials and (b) a single production test trial. Sentences in quotes represent the sentences children heard on that particular trial.

After exposure, we used a wug-style production test to determine whether children had formed a productive rule (Berko, 1958). During this test, we presented participants with a singular image of a noun they had not seen in their exposure and the corresponding singular sentence. Then we asked: "if this one is said *gentif [novel noun]*, how would you say this one?" and showed them a plural image of the same novel noun (see Figure 2b). Each participant completed 12 production test trials, 2 for each of 6 novel nouns (*bleggin, daffin, norg, sep, fluggit*, and *geed*). To prevent participants from using a plural form based on precise number (e.g., add "ka" when there are 4), the test items contained 3 or 5 instances of the novel noun (during exposure, plural sentences were paired with 2, 4, or 6).

Following the production test, we gave participants a rating test to ensure they had learned the nouns and markers they were exposed to. To keep the rating test the same for all participants, we selected four nouns that were marked in the same way in both conditions: two that were marked with the regular form and two that were exceptions. Because the most frequent nouns were marked with the regular form in both conditions, this corresponded to two high-frequency nouns (rank 1 and 2 in the Zipfian distribution) and two low-frequency nouns (rank 7 and 9 in the Zipfian distribution). Each of these four nouns appeared four times in the rating test - twice with the correct marker and twice with an incorrect marker - resulting in 16 total rating test trials. On each trial, we presented participants with a picture of one of the four nouns and its corresponding sentence, with either the correct marker or an incorrect marker. We asked participants to decide, on a scale from 1 to 5, how well the sentence matched the card, where 5 was "definitely matches" and 1 was "definitely does not match".

Predictions

We designed our language such that if children follow the Tolerance Principle, only children in the 5 regular 4 exception condition should form a productive rule. But how will we determine whether a child has formed a productive rule? Prior to running participants, we determined that children who form a productive rule should apply the rule on every single test trial — that is, to 100% of novel nouns. This might seem extreme, particularly for a study involving children, but it is what we expect from productive rules in natural language: they apply to all cases for which there are no known exceptions. Alternatively, if children have not formed a productive rule, we similarly expect them to behave as they do in natural language: never applying the rule, instead producing a bare stem or producing the regular form no more than would be predicted by chance. These predictions thus anticipate strongly categorical behavior from participants who follow the Tolerance Principle.

2.2 Results and discussion

To test whether children formed productive rules in our artificial language, we compared how often children used the regular form on the production test to the values we predicted a priori. Children who formed a productive rule should apply the regular form categorically, on 100% of test trials, while children who did not form a productive rule should apply the regular form no more than is expected by chance. Since learners in the 3R6E condition are exposed to seven markers (the regular form plus six exceptions), chance-level is 1 in 7 (14.29%). Alternatively, if children simply regularize the most frequent token in their input, they should apply the regular form to 100% of test trials in both conditions.

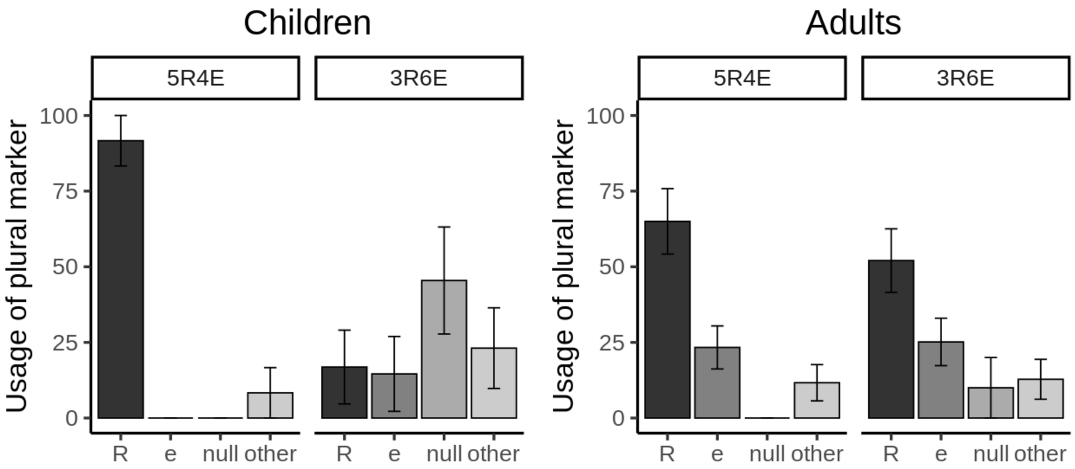


Figure 2. Usage of plural markers at test in the 5 regular, 4 exception (5R4E) condition and 3 regular, 6 exceptions (3R6E) for children and adults in Experiment 1. *R* is the regular form, e is any exception, *null* is no marker, and *other* is a marker not part of the artificial language (e.g. the English plural marker -s). Dashed line is the level of the regular form in the input.

Focusing first on the child data, children in the 5R4E condition were significantly more likely to produce the regular form than children in the 3R6E (logistic mixed effect regression, Est. = 18.72, SE = 5.815, p=0.001), with 5R4E children producing the regular form on 91.67% of production test trials and 3R6E children producing the regular form on only 16.87% of trials. To determine whether these results aligned with our predictions, we next conducted t-tests against our hypothesized values: 100% for rule formation and 14.29% for no rule (Table 2). As we predicted, children in the 5R4E condition nearly always used the regular form (usage was not statistically different from 100%, t(6)=1.00, p=0.18), while children in the 3R6E condition used the regular form no more than predicted by chance (t(7)=0.21, p=0.420).

Children thus appear to form a productive rule when the Tolerance Principle predicts that they will (in the 5R4E condition), but not when it predicts that they won't (in the 3R6E condition). Further, as shown in Figure 3, their behavior is nearly categorical. In the 5R4E condition, all but one child used the regular form on 100% of production trials. In the 3R6E condition, all but one child produced the regular form no more than expected by chance.

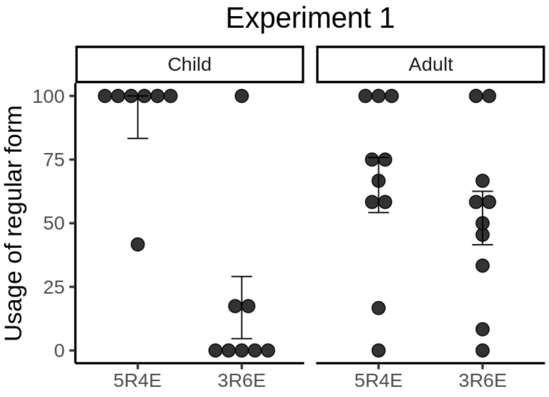


Figure 3. Usage of the regular form by individual children and adults in Experiment 1. Error bars are standard error of the mean.

Unlike children, adults used the regular form significantly less than 100% of the time in the 5R4E condition (m=65.00%, t(9)=-3.24, p<0.01) and significantly more than expected by chance in 3R6E condition (m=52.05%, t(9)=3.60, p<0.01). A logistic mixed effect regression revealed that adults in the 5R4E condition were not more likely to use regular form than adults in the 3R6E condition (Est. = 0.89, SE = 1.09, p =0.42), suggesting that adult behavior does not differ in a significant way across conditions. Why do adults not differ across conditions? One possibility is that they simply produce the regular form with the same frequency it occurred in the input, a behavior known as probability matching, which would lead to approximately the same frequency for the regular form in the two conditions (Austin, Schuler et al., under review; Hudson Kam & Newport, 2005, 2009). Figure 4 shows the usage of the regular form by children and adults alongside the token frequencies of the regular form in the input. Recall that we designed our language such that the regular form was frequent in both conditions: 75% of the plural exposure sentences in the 5R4E condition, and 58.3% of plural exposure sentences in the 3R6E condition. Aligned with our prediction, a t-test revealed that adult productions match the token frequency in both the 5R4E (t(9)=-0.97, p=0.37) and 3R6E conditions (t(9)=-0.68, p=0.51).

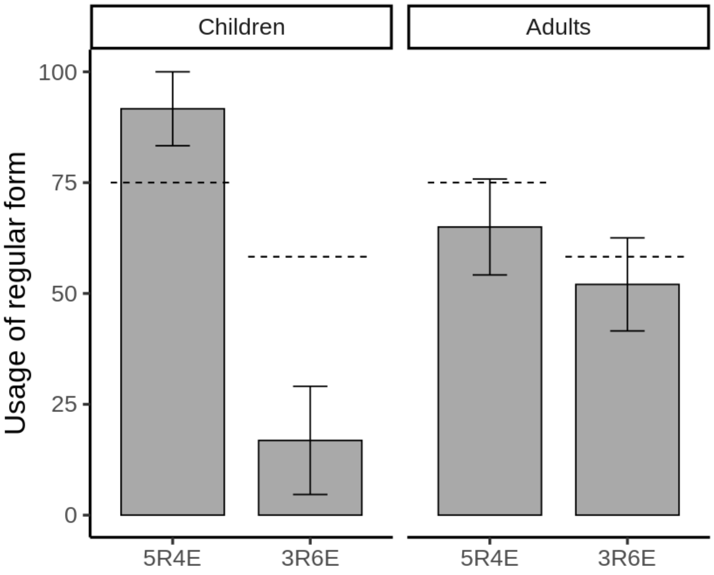


Figure 4. Percentage of regular inflection *ka* applied to novel nouns by children and adults when their exposure contained 5 regulars/4 exceptions compared with 3 regulars/6 exceptions. Dashed line indicates the token frequency of the *ka* inflection in the input.

To summarize, we found that children's generalization behavior was well predicted by the Tolerance Principle. In our category of 9 nouns, children formed a productive generalization when there were 4 exceptions to regular form, but not when there were 6. Moreover, children behaved categorically:

they either formed a productive generalization (applying the regular form to all novel cases), or used the regular form no more than was predicted by chance at test. Adult behavior, on the other hand, was not well predicted by the Tolerance Principle. Adults did not appear to form a productive generalization, instead producing the regular form at approximately the token frequency it occurred in the language input.

3. Experiment 2

As predicted by the Tolerance Principle, in Experiment 1 we found that children formed productive rules based on the number of types that take the regular form, not based on the most frequent tokens in the input. In this experiment we applied the regular form to the most frequent nouns in our exposure corpus, such that the regular form was the most frequent token in both conditions. However, in natural languages, the regular form is not always applied to the most frequent items (as in our Experiment 1), nor are exceptions always the most frequent (though this is the case for some irregular past tense forms in English). Instead, for many patterns, high frequency items can be either regular or irregular: the regular form is often strongly attested at the top end of the frequency rank with exceptions well-distributed among them. For example, while 86% of the 1000 most frequent verbs in child-directed English take the regular past-tense form, 54 of the top 100 are irregular (Pinker, 1999; Yang 2016). For English plural, the majority of the top 100 nouns take the regular form, but there are 6 exceptions distributed among these high-frequency nouns (e.g., people, children, women) (Yang, 2016).

To determine whether the Tolerance Principle predicts when children form productive rules under more natural conditions, we repeated Experiment 1 with a modified (more ecologically valid) artificial language. In the new language, we altered which plural markers applied to which nouns in the frequency rank, so that the regular marker and the exceptions were distributed more evenly across nouns of differing frequencies. This distribution more accurately approximates what we see in natural languages: a mix of regulars and exceptions at the both the high and low end of the frequency distribution.

Will children continue to follow the Tolerance Principle when the regular form does not always apply to the most frequent items in the language? Yang (2016) has tested the Tolerance Principle on dozens of patterns in natural language, including those in which exceptions are highly frequent (e.g., the infamous English past tense) and those for which the rule structure is highly complex (e.g., German plural). We therefore expect children in Experiment 2 to continue to follow the Tolerance Principle, even though the regular form no longer applies to the most frequent nouns. As in Experiment 1, we include a group of adult controls in order to observe whether children and adults behave the same or differently in their acquisition of productive rules.

3.1 Method

Participants

Participants were 20 children (mean = 6.66 years, sd = 0.85 years, range = 5.16 – 7.75 years) and 15 adults (mean = 20.84 years, sd = 2.58 years, range = 18.00 – 25.42 years). We excluded an additional four children and six adults from analysis for failure to meet the experiment inclusion criteria (2 children and 1 adult were non-native English speakers), failure to produce the correct noun on at least 50% of the test trials (2 children), or equipment malfunction (5 adults). Children were recruited from schools, camps, and day-cares in the Washington D.C. metro area and adults were recruited from the Georgetown University community.

All participants were native English speakers with normal to corrected-to-normal hearing and vision. Children received stickers and a set of small toys for their participation and their caregivers received a $10 travel reimbursement if they traveled to our lab. Adults were paid $10 for their participation.

Stimuli

For Experiment 2, we made two changes to the artificial language from in Experiment 1 (described in Section 2.1). First, to more closely mirror natural languages, rather than applying the regular form to the most frequent nouns in both conditions, we distributed the choice of which nouns were regular and which nouns were exceptions evenly along the Zipfian distribution. Second, to ensure that our results would not be driven by the plural marker assigned to the most frequent noun, we created two languages: one in which the most frequent noun took the regular form (Language A) and one in which the most frequent noun was an exception (Language B) (see Table 3).

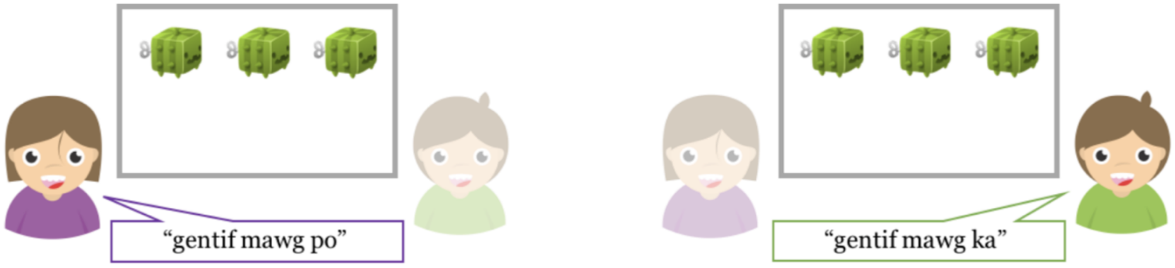
Table 3. Frequency of nouns and their plural markers in Experiment 2, Languages A and B. The regular form appears in bold.

Rank	Freq	N Plural	Noun	Language A		Language B	
				5R4E	*3R6E*	*5R4E*	*3R6E*
1	24	16	mawg	**ka**	**ka**	po	po
2	12	8	tombur	po	po	**ka**	**ka**
3	8	5	glim	**ka**	**ka**	lee	lee
4	6	4	zup	lee	lee	**ka**	**ka**
5	6	4	spad	**ka**	**ka**	bae	bae
6	4	3	daygin	bae	bae	**ka**	**ka**
7	4	3	flairb	**ka**	tay	tay	tay
8	4	3	klidam	tay	muy	**ka**	muy
9	4	3	lepal	**ka**	woo	**ka**	woo

Procedure

The exposure and production test were performed as in Experiment 1 (see section 2.1), except adults in Experiment 2 were run in the lab just like children, not on Mechanical Turk.

In order to collect rating data for every noun (rather than sampling only four of the nouns as in Experiment 1), we created a new two-alternative forced-choice test in which we asked about every noun three times. On each of the 27 trials, participants saw a plural image flanked by two cartoon characters, one dressed in green and the other dressed in purple. The characters took turns describing the image, with one using the correct plural marker and the other using an incorrect plural marker. After hearing both alternatives, we asked the participant to choose the character who "said it best" (see Figure 4). We counterbalanced who spoke first and who said the correct answer across trials.



(a) Incorrect alternative

(b) Correct alternative

Figure 4. Example trial for the two-alternative forced-choice test in Experiment 2. The correct alternative is shown in (b) and the incorrect alternative is shown in (a) for Language A.

3.2 Results & Discussion

First, we asked whether children and adults follow the Tolerance Principle by comparing their usage of the regular form to our predicted values. As in Experiment 1, we predicted that participants who form a productive rule should use the regular form on 100% of production test trials, while participants who have not formed a productive rule should use the regular form no more than predicted by chance (14.29%).

The results are shown in Figure 5. Considering the adult data first, adults in Experiment 2 (as in Experiment 1) do not appear to follow the Tolerance Principle. Adults mark production test trials with the regular form significantly less than 100% of the time in the 5R4E condition (t(7)=--3.27, p<0.001) and significantly more than expected by chance in the 3R6E condition (t(6)=12.18, p<0.001). As in Experiment 1, adults in both conditions appear to match the token frequency of the regular form in the language they were exposed to.

Turning next to the children, in the 3R6E condition children did behave as predicted by the Tolerance Principle, using the regular form no more than one would expect by chance (mean = 7.14%, t(6)=-1.00, p=0.82). However, in the 5R4E condition, children used the regular form significantly less than 100% of the time (mean = 54.63, t(12)=-3.65, p<0.001), our a-priori prediction for productive rule formation. That is, at least in terms of the mean usage of the regular form, these data (shown in Figure 5) suggest that usage of the regular form in the 5R4E condition is not consistent with the prediction of the Tolerance Principle.

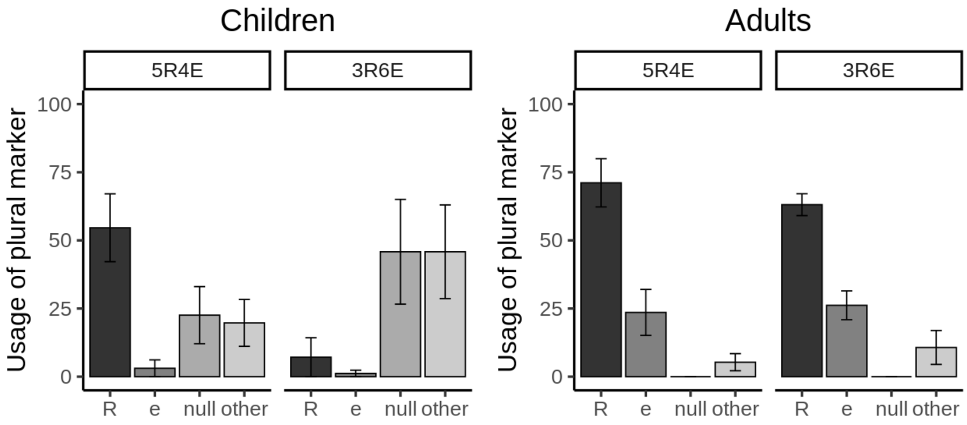


Figure 5. Usage of plural markers at test in the 5 regular, 4 exception (5R4E) condition and 3 regular, 6 exceptions (3R6E) for children and adults in Experiment 2. *R* is the regular form, e is any exception, *null* is no marker, and *other* is a marker not part of the artificial language (e.g., the English plural marker -s).

However, the high standard deviation (46.79%) for the usage of the regular form in the 5R4E condition suggests that a more careful look at the individual children in our experiment may be needed. When we inspected the individual data (shown in Figure 6), the strikingly categorical nature of most children's behavior was apparent. Of the 20 children who participated in this experiment, 16 displayed categorical behavior, using the regular form either 100% of the time (5 children) or no more than expected by chance (11 children). Very few children scored in between, at the mean level of usage for the condition.

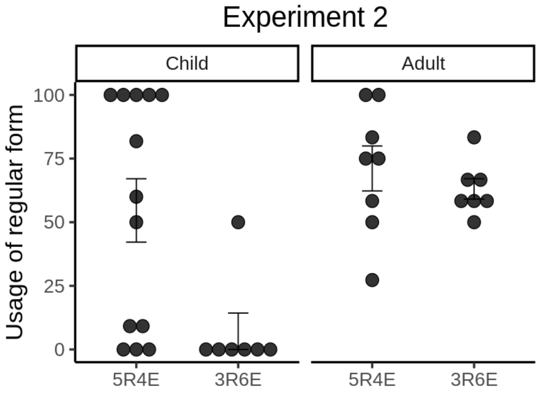


Figure 6. Usage of the regular form by individual children and adults in Experiments 2. Error bars are standard error of the mean.

What was causing this categorical split in the 5R4E condition? In particular, what explains the behavior of the 5 children in the 5R4E condition who definitively do not form a productive rule in this condition, even though the Tolerance Principle predicts that they should?

One possibility is that children in Experiment 2 did not learn all of the noun-marker pairings during exposure. Recall that the number of items in the category, N, is crucial for calculating the threshold for productivity under the Tolerance Principle ($\theta_N \leq N/\ln(N)$). In our 9-noun artificial language, children are predicted to form a productive rule when there are fewer than 4.098 exceptions to the rule. But if a child learned only 7 of the 9 noun-marker pairings we exposed them to, the Tolerance Principle predicts a different threshold for productivity, tolerating only 3.60 exceptions rather than tolerating 4.098 exceptions. Further, if 4 of the 7 nouns the child learned were exceptions, the child would have exceeded the number of exceptions a productive rule should tolerate and would therefore be predicted not to form a rule. This variation in how many nouns are learned is especially likely in Experiment 2, where the crucial nouns carrying the regular marker can be presented infrequently and for this reason may not be well learned.

To test this possibility, we conducted a separate analysis focusing on the children in both conditions who behaved categorically. Here, we used the rating test data to determine which nouns each child learned and calculated a new "personal Tolerance Principle" for each child. To assess which nouns each child learned, we set a criterion for learning a noun: a child was considered to "know" a noun if they provided the correct response on all 3 of the rating test trials. Using this criterion, we counted the number of nouns that each child learned and used this value as the N with which to compute the child’s personal Tolerance Principle threshold ($\theta_N \leq N/\ln(N)$). We next determined how many of each child’s known nouns took the regular form and how many were exceptions. Finally, we compared the number of exceptions each child learned to their personal threshold for productivity, θ_N: the number of exceptions the child should tolerate according to their personal Tolerance Principle threshold.

To illustrate, we can perform these calculations for one of the children in our experiment. Following our noun learning criterion (rating 3/3 trials correct), this child was found to know 8 of the 9 noun-marker pairings presented during exposure (or N = 8). Thus his personal Tolerance Principle threshold (θ_N) was $8/\ln(8) = 3.85$. According to the Tolerance Principle, then, this child should tolerate 3.85 exceptions before forming a productive rule is no longer the most efficient strategy (not 4.098 as originally calculated for our category of 9 nouns). Of the 8 nouns this child learned, 4 were exceptions. Based on his personal Tolerance Principle, the language has exceeded his threshold for productivity and this child is predicted not to form a productive rule.

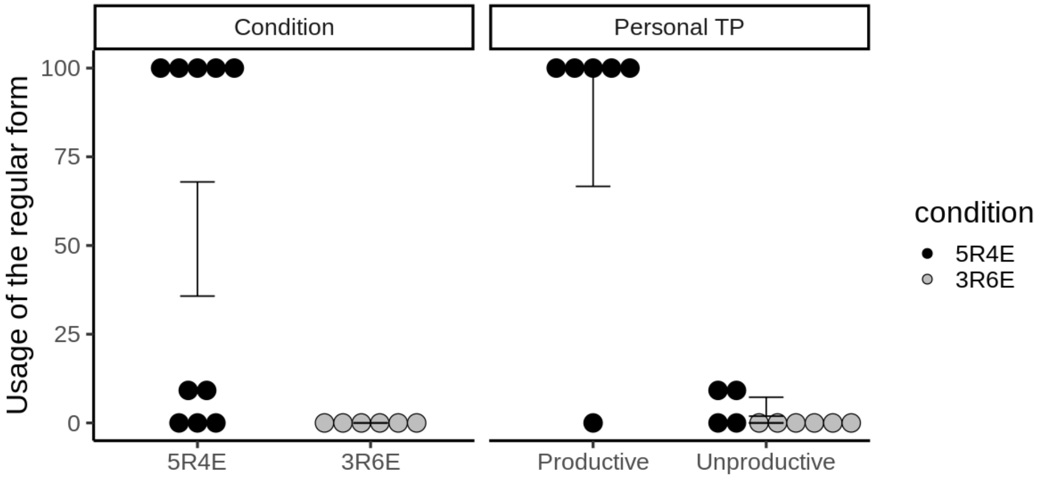


Figure 7. Usage of the regular form by categorical children in Experiment 2, visualized by condition, personal Tolerance Principle, and Language. Error bars are standard error of the mean.

In Figure 7, the individual children's data are re-plotted according to the predictions of their personal Tolerance Principle (as "Productive" or "Not Productive"), compared to the original 5R4E and 3R6E conditions. Specifically, of the 5 children who did not form a productive rule in the 5R4E condition, 4 had a personal Tolerance Principle below the threshold for productivity. The resulting data overall now match very closely with what the Tolerance Principle predicts: 15 of the 16 children use the regular form as predicted by their personal Tolerance Principle.

4. General discussion

In Experiment 1, we asked whether the Tolerance Principle accurately predicted when children would and would not form a productive rule in an artificial language learning experiment. For a category of 9 nouns, the Tolerance Principle predicts that a productive rule will be formed when there are fewer than 4.096 exceptions. We found that children formed productive rules just as the Tolerance Principle predicted:

when there were 4, but not 6, exceptions to the rule. Importantly, our analysis asked whether learners extended the rule to 100% of the test trials — the most rigorous possible test of productivity. We found that, while both children and adults were more likely to extend a productive rule when there were 4 exceptions than when there were 6, only children displayed a categorical distinction between forming a productive rule and not forming one. As predicted by the Tolerance Principle, almost every child exposed to 5 regulars and 4 exceptions extended the rule to 100% of test trials, while almost no children exposed to 3 regulars and 6 exceptions did (see Fig 3). While children exposed to 6 exceptions did occasionally use the regular form on novel items, their results are only what one would expect under conditions of uncertainty: children either used no plural marker at all (omission) or otherwise selected at random from among the various markers they heard during exposure.

In Experiment 2, we asked whether the Tolerance Principle would continue to predict the conditions for productive rule formation when the regular and exceptional forms were distributed in a way more similar to their frequencies in natural languages. While our initial analysis did not suggest that children's behavior was in accord with the Tolerance Principle, the individual data revealed that children did behave remarkably categorically and suggested that the mean performance did not accurately reflect that of the individuals'. Children either formed a productive rule or did not, just as they did in Experiment 1. To explain these categorical results, we included a rating test in Experiment 2 that allowed us to make an estimate of each child's individual vocabulary size in the artificial language. When taking into account the number of lexical items that each child had learned (the true N in the Tolerance Principle calculations), children's behavior in Experiment 2 was indeed well-predicted by the Tolerance Principle.

Taken together, the findings from these experiments illustrate the power of models like the Tolerance Principle. By making clear predictions about when productive generalizations should and should not be made, we can test the model and even take into account the individual child's input and successful learning to determine whether a generalization will be made.

Why does the Tolerance Principle work for children and not adults?

While our primary goal was to ask whether the Tolerance Principle accurately predicts when children will and will not form productive rules, a second finding emerged. Though the behavior of children in our experiments was well predicted by the Tolerance Principle, the behavior of adults was not. Instead, adult productions in both experiments closely matched the token frequency of the regular form in the input, indicating that adults engage a learning mechanism that makes use of this different statistic. This finding is not necessarily surprising. Adults have been shown to engage in this behavior — often called probability matching — in many diverse learning tasks (Neimark 1956; Gardner 1957; Gardner 1958; Weir 1972), including during language learning (Austin, Schuler et al., under review; Hudson Kam and Newport 2005; Hudson Kam and Newport 2009). Probability matching is a behavior in which learners match the probability with which they have observed or encountered a stimulus. For example, when asked to guess which of two lightbulbs will illuminate on a given trial, adults guess in line with observed probabilities: if the left lightbulb has been correct 75% of the time and the right 25%, they'll guess left on 75% of the test trials and guess right on 25% of the test trials (Gardner 1957; Weir 1964; Weir 1972). Similarly, in other artificial language learning studies, adults' productions match the probability with which a given form occurs in their input (Austin, Schuler et al., under review; Hudson Kam and Newport 2005; Hudson Kam and Newport 2009): if one morpheme occurs in 67% of the occurrences of a particular context and another morpheme occurs in 33% of the occurrences of that context, adults will produce these forms at these same probabilities.

Notably, this probability matching behavior is not the optimal strategy under these circumstances. In the light bulb studies, learners who use a strategy called maximizing — always guessing the highest probability lightbulb — would achieve the highest correct rate. Interestingly, children, but not adults, are widely known to engage in this maximization behavior, both in the probability learning literature (Stevenson and Weir 1959; Weir 1964; Derks and Paclisanu 1967) and during language learning (here called regularization) (Austin, Schuler et al., under review; Hudson Kam and Newport 2005; Hudson Kam and Newport 2009). In the context of productivity in language, probability matching is also arguably not the optimal strategy. Given the nature of natural language rules, successful language acquisition requires productive generalizations. Learners hear a finite sample of the possible utterances in the language and, from that sample, they must generate an infinite number of novel grammatical utterances, even those they have never encountered before. Matching the probability with which particular forms are used does not achieve this end result with the same success that forming appropriate productive generalizations does.

One question worth exploring, then, is why adults engage in probability matching when doing so is suboptimal in these contexts. One possibility is that children are in a maturational state that is optimized for forming simple, clean, productive generalizations, while adults are optimized for something else. Said another way, children are better at acquiring language because of (not in spite of) their cognitive limitations compared to adults (Newport 1990). There are a number of ways in which children and adults differ — particularly ways in which children are more cognitively limited than adults — that lend support to this explanation.

For example, Thomson-Schill et al. (2009) argue that one way children are more cognitively limited than adults is in the maturation of the prefrontal cortex (PFC). The PFC takes a long time to develop in humans (e.g. Rakic et al. 1986; Chugani and Phelps 1986; Huttenlocher and Dabholkar 1997), meaning that children are generally much worse than adults at tasks requiring cognitive control (Diamond and Doar 1989). Performance on the Stroop task, for example, requires participants to suppress the meaning of a color word in order to name the correct ink color (e.g., the word "red" printed in green ink) — adults are significantly better at this than children (MacLeod 1991; Demetriou et al. 2001; Adleman et al. 2002; Hanauer and Brooks 2003). However, Thompson-Schill et al (2009) have proposed that these same cognitive control abilities may actually interfere during learning processes that depend on low-level competition mechanisms. In other words, a mature PFC may allow adults to override low-level competition mechanisms to make an informed guess. Perhaps the Tolerance Principle is part of a low-level competition mechanism with which productive generalizations can be formed. Under this account, adult behavior in our experiment is not well captured by the Tolerance Principle because their more fully developed cognitive control abilities allow them to override this mechanism and make (what they believe to be) an informed guess: I heard the "ka" inflection about 65% of the time, so I'll guess "ka" 65% of the time.

Another way in which children's cognitive limitations may confer a language learning advantage comes from the development of memory systems. In the declarative-procedural memory framework, irregular forms are thought to be stored in declarative memory while the regular form is thought to be composed in real-time by the procedural memory system (Ullman et al. 1997; Ullman 2001). Interestingly, a large body of evidence suggests that the procedural memory system is functioning at its peak during childhood, then progressively declines from adolescence through early adulthood (Fredriksson et al. 2000;

Schlaug 2001; Walton et al. 1992; Wolansky et al. 1999; Janacsek et al. 2012), while the declarative memory system shows the opposite developmental pattern (Campbell and Spear 1972; DiGiulio et al. 1994; Kail and Hagen 1977; Meudell 1983; Ornstein 1978; Siegler 1978; Finn et al. 2016). That is, children's memory systems are in a maturational state that favors the formation of productive rules over storing individual forms. In support of this account, recent research investigating how gray matter volume changes across development has found that the volume of basal ganglia structures (thought to underlie procedural memory) decreases with age while the volume of the hippocampus (thought to underlie declarative memory) shows an inverted u-shaped trajectory (Wierenga et al. 2014).

Finally, Newport's Less is More hypothesis proposes that another advantage children have over adults is that they "start small," extracting from the speech stream and operating over smaller and more fundamental units of language (Newport, 1990; see also Elman 1993). Interestingly for the Tolerance Principle, the concept of starting with a small number of elements results in a quantifiable computational advantage (see Yang, 2016); that is, the proportion of exceptions a productive rule can tolerate dramatically decreases as the number of items to which the rule may apply (N), increases. Yang (2016) suggests that this implies "smaller is better" for the acquisition of productive rules. Thus, children with their smaller vocabularies will have an easier time acquiring productive rules than adults.

While this likely contributes to the natural language acquisition advantage enjoyed by children, our results suggest that this cannot be the whole story. In the experiments presented in this paper, both children and adults arguably "started small", as the artificial languages they were exposed to contained only a small number of lexical items. Both age groups had a domain size of 9 (N=9), yet only children follow the Tolerance Principle in our experiments. This suggests that there must be additional sources of difference between children and adults that could explain why the Tolerance Principle is unique to children here. Perhaps some combination of all of the accounts above (and others as well) that creates an optimal maturational state in which to acquire a productive rule.

The utility of models like the Tolerance Principle

In the present experiments we have argued that the Tolerance Principle appears to capture something fundamental about the way in which children form productive generalizations. While more work is needed to determine precisely what this fundamental learning process is, our experiments demonstrate the utility of investigating generalization behavior with the Tolerance Principle and models like it. As we've pointed out elsewhere, the Tolerance Principle has positioned itself in an extremely useful (and previously nearly vacant) space in the productivity literature: it makes a clear prediction about which processes should be productive and which should not. And unlike previous mathematical and computational models of learning and generalization (e.g., Anderson 1990, Nosofsky et al. 1994, Tenenbaum & Griffiths, 2001), there are not free parameters to fit: the vocabulary composition of the child learner yields a unique prediction. Needless to say, the Tolerance Principle must be tested on many more additional empirical cases. However, artificial language learning tasks such as the current study can provide precise measures over the learner's vocabulary and experience that are near impossible to provide in naturalistic studies.

5. Conclusion

In the present work, we set out to determine whether the Tolerance Principle (Yang, 2016), a recent model shown to predict when productive generalizations will be formed in natural language, would predict children's generalization behavior in an artificial language learning experiment with similar precision. Across two experiments we found that children's behavior was well predicted by the Tolerance Principle, not the token frequency of the regular form (Experiment 1), and that the Tolerance Principle could be applied to children's word learning on an individual level to predict when they will or will not generalize to novel forms (Experiment 2). We argue that the literature on productive generalization has, until now, lacked a formal model that makes a precise prediction about when generalizations will be formed. We have explored why this model might more accurately predict children's behavior in this task than adults and pointed out that it offers a positive-evidence approach to acquiring productive generalizations that is nearly unique in the language acquisition literature. Although future research is certainly needed, we believe these experiments contribute to our understanding of the mechanisms of productive generalization and to some of the advantages that children may hold over adults as they learn language.

References

Adleman, N. E., Menon, V., Blasey, C. M., White, C. D., Warsofsky, I. S., Glover, G. H., & Reiss, A. L. (2002). A developmental fMRI study of the Stroop color-word task. *Neuroimage*, *16*(1), 61–75.

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

Allen, S. (1996). *Aspects of argument structure acquisition in Inuktitut* (Vol. 13). John Benjamins Publishing.

Ambridge, B. (2010). Children's judgments of regular and irregular novel past-tense forms: New data on the English past-tense debate. *Developmental Psychology*, *46*(6), 1497.

Ambridge, B., Pine, J. M., Rowland, C. F., & Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, *106*(1), 87-129.

Anderson, S. R. (1969). *West Scandinavian vowel systems and the ordering of phonological rules*. Massachusetts Institute of Technology.

Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.

Aronoff, M. (1976). *Word formation in generative grammar*. MIT Press.

Austin, Schuler, Furlong, & Newport (under review). Learning a language from inconsistent input: Regularization in child and adult learners.

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: a corpus-based study. *Linguistics and Philosophy*, *29*(5), 801–844.

Baayen, R. H., & Renouf, A. (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, 69–96.

Baerman, M., Corbett, G., & Brown, D. (2010). Defective paradigms: Missing forms and what they tell us.

Berko, J. (1958). The child's learning of English morphology. *Word & World*, *14*(2-3), 150–177.

Björnsdóttir, S. M. (2021). Productivity and the acquisition of gender. Journal of Child Language. 1-26.

Bloch, B. (1947). English verb inflection. *Language*, *23*(4), 399-418.

Brown, D., & Hippisley, A. (2012). *Network morphology: A defaults-based theory of word structure* (Vol. 133). Cambridge University Press.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*(5), 425–455.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form* (Vol. 9). John Benjamins Publishing.

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

Bybee, J. L., & Moder, C. L. (1983). Morphological classes as natural categories. *Language*, 251–270.

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

Campbell, B. A., & Spear, N. E. (1972). Ontogeny of memory. *Psychological Review*, *79*(3), 215.

Caprin, C., & Guasti, M. T. (2009). The acquisition of morphosyntax in Italian: A cross-sectional study. *Applied Psycholinguistics*, *30*(01), 23–52.

Cazden, C. B. (1968). The acquisition of noun and verb inflections. *Child Development*, 433–448.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Chugani, H. T., & Phelps, M. E. (1986). Maturational changes in cerebral function in infants determined by FDG positron emission tomography. *Science*, *231*, 840–844.

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *The Behavioral and Brain Sciences*, *22*(06), 991–1013.

Clahsen, H., & Penke, M. (1992). The acquisition of agreement morphology and its syntactic consequences: New evidence on German child language from the Simone-Corpus. In *The acquisition of verb placement* (pp. 181–223). Springer.

Corbett, G., Hippisley, A., Brown, D., & Marriott, P. (2001). *In frequency and the emergence of linguistic structure* (pp. 201–226). Amsterdam: John Benjamins.

Dabrowska, E. (2001). Learning a morphological system without a default: The Polish genitive. *Journal of child language*, *28*(3), 545-574.

de Vries, H., Meyer, C., & Peeters-Podgaevskaja, A. (2020). Learning strategies in Russian ordinal acquisition. First Language

Deen, K. U. (2005). *The acquisition of Swahili* (Vol. 40). John Benjamins Publishing.

Demetriou, A., Spanoudis, G., Christou, C., & Platsidou, M. (2001). Modeling the Stroop phenomenon: Processes, processing flow, and development. *Cognitive Development*, *16*(4), 987–1005.

Demuth, K., & Nurse, D. (2003). The acquisition of Bantu languages. *The Bantu Languages*, 209–222.

Derks, P. L., & Paclisanu, M. I. (1967). SIMPLE STRATEGIES IN BINARY PREDICTION BY CHILDREN AND ADULTS. *Journal of Experimental Psychology*, *73*(2), 278.

Diamond, A., & Doar, B. (1989). The performance of human infants on a measure of frontal cortex function, the delayed response task. *Developmental Psychobiology*, *22*(3), 271–294.

DiGiulio, D. V., Seidenberg, M., Oleary, D. S., & Raz, N. (1994). Procedural and declarative memory: A developmental study. *Brain and Cognition*, *25*(1), 79–91.

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

Fernändez-Dobao, A., & Herschensohn, J. (2020). Present tense verb morphology of Spanish hl and l2 children in dual immersion: Feature reassembly revisited. Linguistic Approaches to Bilingualism, 10(6), 775–804.

Finn, A. S., Kalra, P. B., Goetz, C., Leonard, J. A., Sheridan, M. A., & Gabrieli, J. D. (2016). Developmental dissociation between the maturation of procedural memory and declarative memory. *Journal of Experimental Child Psychology*, *142*, 212–220.

Fredriksson, A., Schröder, N., Eriksson, P., Izquierdo, I., & Archer, T. (2000). Maze learning and motor activity deficits in adult mice induced by iron exposure during a critical postnatal period. *Developmental Brain Research*, *119*(1), 65–74.

Garcia, G. D. (2019). When lexical statistics and the grammar conflict: learning and repairing weight effects on stress. Language, 95(4), 612–641.

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

Gardner, R. A. (1958). Multiple-choice decision-behavior. *The American Journal of Psychology*, *71*(4), 710–717.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2016). Partial productivity of linguistic constructions: Dynamic categorization and statistical preemption. *Language and Cognition*, *8*(3), 369–390.

Guion, S. G. (2003). The vowel systems of Quichua-Spanish bilinguals. *Phonetica*, *60*(2), 98-128.

Gxilishe, S., de Villiers, P., de Villiers, J., Belikova, A., Meroni, L., & Umeda, M. (2007). The acquisition of subject agreement in Xhosa. *Proceedings of the Conference on Generative Approaches to Language Acquisition (GALANA)*, *2*, 114–123.

Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic inquiry*, *4*(1), 3-16.

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *The view from building 20* (pp. 111–176). The MIT Press.

Halle, M., & Mohanan, K. P. (1985). Segmental phonology of modern English. *Linguistic Inquiry* , *16*(1), 57–116.

Hanauer, J. B., & Brooks, P. J. (2003). Developmental change in the cross-modal Stroop effect. *Perception & Psychophysics*, *65*(3), 359–366.

Hoeffner, J. H. (1997). Are rules a thing of the past? A single mechanism account of English past tense acquisition and processing. *Unpublished Ph. D. Dissertation. Pittsburgh, PA: Carnegie Mellon University*.

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: The Official Journal of the Society for Language Development*, *1*(2), 151–195.

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

Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology*, *387*(2), 167–178.

Ivimey, G. P. (1975). The development of English morphology: an acquisition model. *Language and speech*, *18*(2), 120-144.

Janacsek, K., Fiser, J., & Nemeth, D. (2012). The best time to acquire new skills: Age-related differences in implicit sequence learning across the human lifespan. *Developmental Science*, *15*(4), 496–505.

Kail, R. V., & Hagen, J. W. (1977). *Perspectives on the development of memory and cognition*. Lawrence Erlbaum Associates Hillsdale, NJ.

Kelly, M. H., & Bock, J. K. (1988). Stress in time. *Journal of experimental psychology: human perception and performance*, *14*(3), 389.

Kiparsky, P. (1982). From Cyclic Phonology to Lexical Phonology. In H. van der Hulst & N. Smith (Eds.), *The Structure of Phonological Representations* (Vol. 1, pp. 131–175). Foris.

Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 589–600.

Labov, W. (2020). The regularity of regular sound change. Language, 96(1), 42–59.

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

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the society for research in child development*, 1-123.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163.

Maratsos, M. (2000). More overregularizations after all: new data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen & Xu. *Journal of Child Language*, *27*(01), 183–212.

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of child language*, *21*(2), 339-366.

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

Maslen, R. J. C., Theakston, A. L., Lieven, E. V. M., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research: JSLHR*, *47*(6), 1319–1333.

Merkuur, A., Don, J., Hoekstra, E., & Versloot, A. P. (2019). Competition in Frisian past participles. In *Competition in inflection and word-formation* (pp. 195–222). Springer.

Meudell, P. (1983). The development and dissolution of memory. *Memory in Animals and Humans*, 83–132.

Mohanan, K. P. (1986). *The theory of lexical phonology* (Vol. 6). D. Reidel Publishing Company.

Neimark, E. D. (1956). Effects of type of nonreinforcement and number of alternative responses in two verbal conditioning situations. *Journal of Experimental Psychology*, *52*(4), 209.

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

Nida, E. (1945). Linguistics and ethnology in translation-problems. *Word*, *1*(2), 194-208.

Nida, E. A. (1949). Morphology: The descriptive analysis of words.

O‘Donnel, T. J. (2011) *Productivity and reuse in language.* Doctoral Dissertation*,* Harvard University.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Ornstein, P. A. (1978). *Memory Development in Children*. Psychology Press.

Pinker, S. (1991). Rules of language. *Science*, *253*(5019), 530.

Pinker, S. (1995). Why the child holded the baby rabbits: A case study in language acquisition. *An Invitation to Cognitive Science*, *1*, 107–133.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books, New York.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, *38*(1), 43–102.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*(1), 1–56.

Pullum, G., & Wilson, D. (1977). Autonomous syntax and the analysis of auxiliaries. *Language*, 741-788.

Rakic, P., Bourgeois, J.-P., Eckenhoff, M. F., Zecevic, N., & Goldman-Rakic, P. S. (1986). Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science*, *232*, 232–236.

Robenalt, C., & Goldberg, A. E. (2015). Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics*, *26*(3), 467-503.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2*, 216–271.

Schlaug, G. (2001). The brain of musicians. *Annals of the New York Academy of Sciences*, *930*(1), 281–299.

Siegler, R. (1978). *Children's thinking: What develops?* Psychology Press.

Sims, A. D. (2006). *Minding the gaps: Inflectional defectiveness in a paradigmatic theory* (Doctoral dissertation, The Ohio State University).

Stefanowitsch, A. (2008). Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics*, *19*(3), 513-531.

Stemberger, J. P. (1983). Inflectional malapropisms: Form-based errors in English morphology. *Linguistics and Philosophy*, *21*(4).

Stevenson, H. W., & Weir, M. W. (1959). Variables affecting children's performance in a probability learning task. *Journal of Experimental Psychology*, *57*(6), 403.

Stump, G. T. (2001). *Inflectional morphology: A theory of paradigm structure* (Vol. 93). Cambridge University Press.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition*, *86*(2), 123-155.

Thompson-Schill, S. L., Ramscar, M., & Chrysikou, E. G. (2009). Cognition without control when a little frontal lobe goes a long way. *Current Directions in Psychological Science*, *18*(5), 259–263.

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. J., & Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, *9*(2), 266–276.

Ullman, M. T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, *30*(1), 37–69.

Walton, K., Lieberman, D., Llinas, A., Begin, M., & Llinas, R. (1992). Identification of a critical period for motor development in neonatal rats. *Neuroscience*, *51*(4), 763–767.

Weir, M. W. (1964). Developmental changes in problem-solving strategies. *Psychological Review*, *71*(6), 473.

Weir, M. W. (1972). Probability performance: Reinforcement procedure and number of alternatives. *The American Journal of Psychology*, 261–270.

Wierenga, L., Langen, M., Ambrosino, S., van Dijk, S., Oranje, B., & Durston, S. (2014). Typical development of basal ganglia, hippocampus, amygdala and cerebellum from age 7 to 24. *NeuroImage*, *96*, 67–72.

Wolansky, M. J., Cabrera, R. J., Ibarra, G. R., Mongiat, L., & Azcurra, J. M. (1999). Exogenous NGF alters a critical motor period in rat striatum. *NeuroReport*, *10*(13), 2705–2709.

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, *22*(03), 531–556.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.

Yang, C. (2005). On productivity. *Linguistic Variation Yearbook*, *5*, 265–302.

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford University Press on Demand.