A User's Guide to the Tolerance Principle

Charles Yang
charles.yang@ling.upenn.edu

Learning a language requires discovering rules that generalize beyond a finite sample of data. The Tolerance Principle (TP) is a theory of how such generalizations are formed. Specifically,

(1) Let a rule R be defined over a set of N items. R is productive if and only if e, the number of items not supporting R, does not exceed θ_N:

$$e \leq \theta_N = \frac{N}{\ln N}$$

If e exceeds θ_N, then the learner will "lexicalize" only these and not generalize beyond them: that is, R is unproductive.

This note provides a summary of the conceptual and methodological issues in the application of the TP, compressing materials published in *The Price of Linguistic Productivity* (Yang 2016) and elsewhere.

1 The Tolerance Principle

The TP builds on the intuition, shared by many (e.g., Aronoff 1976, Plunkett and Marchman 1993, Bybee 1995), that rules must "earn" productivity by the virtue of being applicable to a sufficiently large number of candidates it is eligible for. If there are 10 examples and all but one (9/10) support a rule, generalization ought to ensue. But no one in their right mind would extend a rule on the basis of 2/10: the learner should just memorize the two supporting examples. Productivity is a calibration of regularities and exceptions — crucially with respective to word *types* rather than tokens.

The design principle behind the TP is that the learner favors a more efficient organization of the grammar, measured in terms of real-time language processing. Linguists have traditionally used the Elsewhere Condition (Anderson 1969, Kiparsky 1973, Aronoff 1976) to describe productive rules with exceptions: the exceptions have to be checked off first before the application of the rule. Somewhat surprisingly, reaction-time studies provide direct support for the Elsewhere Condition as a psychological processing model. The order in which the irregulars are listed is frequency sensitive and more importantly, irregulars are processed faster than regulars because they are handled earlier in the search process; see Yang (2016, Chapter 3) for important details. These motivations allow us to establish a cost/benefit calculus for productivity. The learner will choose the faster grammar of the two: (a) a productive rule preceded by a list of exceptions, or (b) no productive rule where every item is lexically listed. The categorical nature of productivity is a matter of necessity for the TP and is strongly supported by the cross-linguistic studies of language acquisition (see Lignos and Yang 2016 for review).

The derivation of the TP assumes that word frequencies follow Zipf's Law (1949), that the rank (r) and frequency (f) of words multiply to a constant (C), or $rf = C$. Thus, a probability p_i of the i-th ranked (out of N) word in a corpus can be expressed as follows:

$$
\begin{aligned}
p_i &= f_i / \sum_{k=1}^{N} f_k \\
&= \Big(\frac{C}{r_i}\Big) / \sum_{k=1}^{N} \frac{C}{r_k} \\
&= \frac{1}{iH_N} \text{ where } H_N = \sum_{k=1}^{N} \frac{1}{k}, \text{ the } N\text{-th Harmonic number}
\end{aligned}
$$

The i-th item on a list takes i units of time in a serial search process after the more frequent, and higher-ranked, items are scanned through. Thus, the expected time complexity to process the no-productive-rule option, i.e., a list of N items all of which are in effect exceptions, is

$$T(N, N) = \sum_{r=1}^{N} r \frac{1}{rH_N} = \frac{N}{H_N}$$

Similar calculation can be made for $T(N, e)$, where a productive rule defined over N words with e exceptions, which must be listed before the application of the rule. Recall that e is the number of exceptions that are ranked by frequency: under the Zipfian assumption, the expected time for accessing the exceptions is $T(e, e)$ or e/H_e. For the other $(N - e)$ items, the access time is the constant e, the number of exceptions. Thus, the overall average for the rule plus exception model is:

(2)

$$
\begin{aligned}
T(N, e) &= \frac{e}{N} T(e, e) + (1 - \frac{e}{N})e \\
&= \frac{e}{N} \frac{e}{H_e} + (1 - \frac{e}{N})e
\end{aligned}
$$

Before we derive an analytical solution to $T(N, N) = T(N, e)$ for the variable e, which will give the critical threshold for productivity, consider the relationship between the two quantities in Figure 1 based on a numerical simulation. The dotted line represents the expected search time for a list of $N = 100$ items, or $T(100, 100)$, which is obviously a constant. The solid line represents the expected search time for having a productive rule with an increasing number of exceptions (e), from 1 all the way up to $N = 100$. Figure 1 shows that when there are few exceptions, i.e. e is small, it is more economical to scan through them before invoking the productive rule. But as e increases, roughly at the value of $e = 21$, it becomes more economical to simply list everything.

Sam Gutmann helped derive a closed form solution to $T(N, N) = T(N, e)$. First, we approximate the harmonic number $H_N = \sum_{i=1}^{N} \frac{1}{i}$ with the natural log ($\ln N$). This approximation only holds when N is fairly large but turns out to be empirically accurate for reasons not understood (Section 4). We would like to find $x = e/N$ such that

$$x \frac{e}{\ln e} + (1 - x)e = \frac{N}{\ln N}$$

Dividing both sides by N and making use of a fact about logarithm:

$$x^2 \frac{1}{\ln N + \ln x} + (1 - x)x = \frac{1}{\ln N}$$

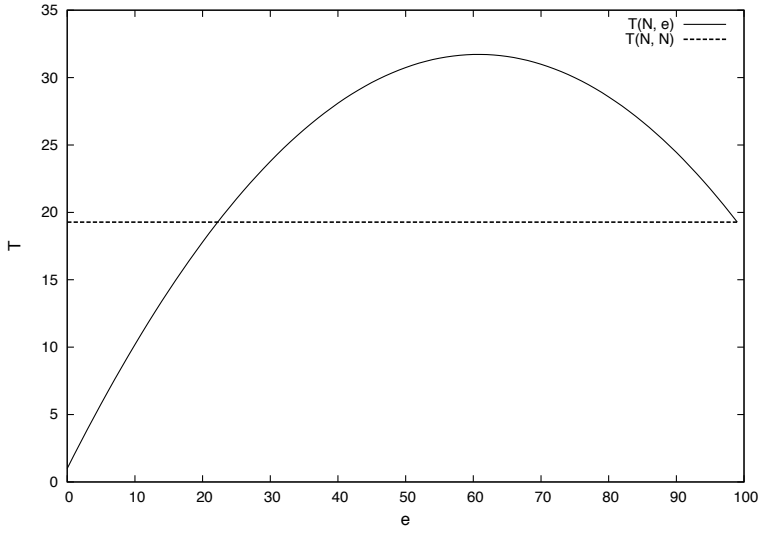

Figure 1: $T(N, N)$ vs. $T(N, e)$, where $N = 100, 1 \leq e \leq 100$.

Let:

$$f(x) = x^2 \frac{1}{\ln N + \ln x} + (1 - x)x - \frac{1}{\ln N}$$

Observe:

$$\begin{aligned} f(\frac{1}{\ln N}) &= \frac{(1/\ln N)^2}{\ln N + \ln \ln N} + (1 - \frac{1}{\ln N})\frac{1}{\ln N} - \frac{1}{\ln N} \\ &= -(\frac{1}{\ln N})^2 + (\frac{1}{\ln N})^3 \frac{\ln N}{\ln N + \ln \ln N} \\ &\approx -\Big(\frac{1}{\ln N}\Big)^2 \\ &\approx 0 \quad \text{for large values of } N \end{aligned}$$

The TP was introduced as a model of learning with attested exceptions: for example, the English irregulars are exceptions to the regular rule by the virtue of not taking *-ed*. Yang (2016, 177) introduces a corollary, the Sufficiency Principle, which specifies how generalizations are formed when the "exceptions" are not attested — but cannot be regarded as impossible, on the ground that evidence of absence is not absence of evidence. But it is clear that both principles generalize on the number of positive examples: specifically, $N - \theta_N$, a super majority of N. Table 1 provides some sample values of N and the associate threshold values θ_N. Note that θ_N decreases quite sharply as a proportion of N, which suggests that rules defined over a smaller vocabulary can tolerate relatively more exceptions, and are thus easier to learn. This has interesting consequences for language development and provides a theoretical underpinning for the idea of "less is more" (Newport 1990, Elman 1993) although I will not dwell on that point here.

In what follows, I discuss two major questions concerning the application of the TP, which lives and dies by the number: the estimation of N and e to calculate productivity, and the nature of the rules and representations under evaluation.

Table 1: The maximum number of exceptions for a productive rule over N items.

N	θ_N	%
10	4	40.0
20	6	30.0
50	12	24.0
100	21	21.0
200	37	18.5
500	80	16.0
1,000	144	14.4
5,000	587	11.7

2 Counting Words

By hypothesis, productivity is determined by two integer values (N and e), which are obviously matters of individual vocabulary variation. Thus the TP allows room for variation in the transient stages of language acquisition as well as in the stable grammars of individual speakers. The relationship between N and e, which may change during the course of language acquisition, determines the status of the rule. If e is very low as a proportion of N, then children may rapidly conclude that a rule is productive. Otherwise, a protracted stage of conservatism may ensue, which may be followed by the sudden onset of productivity, as can be seen, famously, in the over-regularization of irregular verbs (Marcus et al. 1992, Yang 2002). It is also possible that no rule ever reaches the productivity threshold; gaps and other phenomena of ineffability arise.

It is thus important to obtain reliable estimates of N and e, ideally at the individual level. This is sometimes possible. First, in studies of rule learning with artificial languages, the items children are exposed to are under the complete control of the researcher. Further tests can be administered to identify exactly the items that individual children have learned in the training phase — not all children will learn all, or the same, set of items. This can lead to individualized calculation of productivity and subsequent verification (Schuler 2017). Second, there are some, not nearly enough, dense longitudinal records of children's production in the public domain (MacWhinney 2000). Yang (2016, Section 4.1) contains a study of several English-learning children's inflectional morphology acquisition: the significant individual differences can be attributed to vocabulary size and trajectories of growth, and ultimately N and e. Finally, we may turn to established methods for vocabulary estimates (e.g., Fenson et al. 1993, 1994, Hart and Risley 1995) which are then used for TP calculations; see, for example, Yang (2016, Section 5.3) for an individual-level study of the variation and change in the productivity of case marking in Icelandic.

In general, however, vocabulary estimation of language learners is difficult. The challenge is even more formidable when a child-directed speech corpus is not available. For example, the TP provides a precise measure of productivity which has immediate applications in the study of language change, but a small collection of historical texts clearly does not represent the input to language learners centuries ago. Yet there are reasons to believe that the data poverty problem is not debilitating for a quantitative approach to productivity.

First, it's important to remind ourselves that children learn languages very early and accurately, and the terminal state of language acquisition across individuals in a speech community is remarkably uniform as shown in detailed studies of language variation (Labov 1972, Labov et al. 2006). As a logical consequence, the rules of language must be learnable with a very small vocabulary of fairly common words. For instance, a three-year-old's knowledge about the word order and inflectional morphology of their native language is

generally perfect: this is age by which even the most fortunate have only just over 1,000 words in their lexicon. Figure 2 provides the vocabulary size estimates of American English-learning children from a large scale study (Hart and Risley 1995).

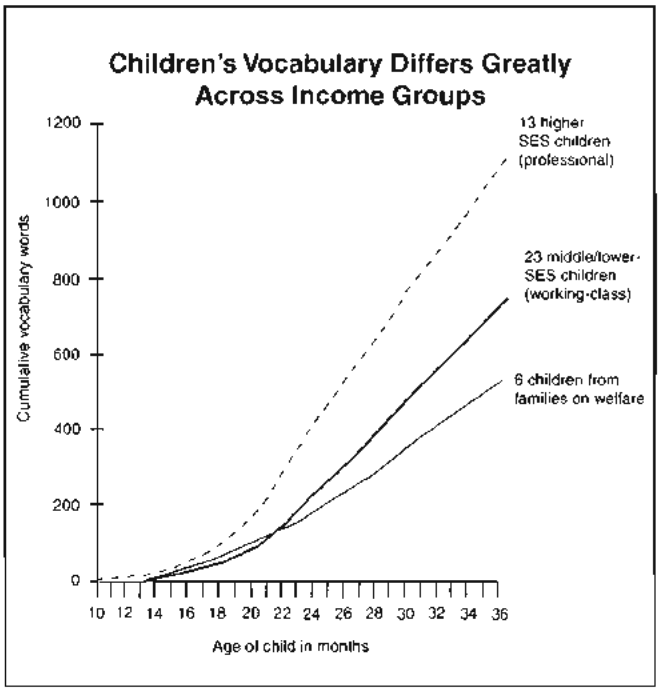

Figure 2: Vocabulary size estimates of American children.

Second, there is converging evidence that lexical frequency can help to provide a reasonable approximation of vocabulary. Nagy and Anderson (1984), for instance, estimate that most English speakers know words above a certain frequency threshold (about once per million). Developmentally, it is also known that children's vocabulary acquisition correlates with word frequencies in child-directed speech (Goodman et al. 2008), especially for open-class words, the primary arena of rule productivity. Here I present an analysis of the Chicago Corpus, a large longitudinal study of vocabulary acquisition by 62 children (Rowe and Goldin-Meadow 2009, Rowe et al. 2012; see the appendix of Carlson et al. 2014). The corpus contains 562 words which have been assessed to be known to English-learning children prior to 50 months. Virtually all these words can be found in the top 1,800 most frequent words in a five-million-word corpus, or roughly a year's worth, of child-directed English extracted from CHILDES — which is broadly consistent with previous vocabulary size studies including Figure 2. Thus, we can hypothesize that a vocabulary size of roughly 2,000 from a relatively large corpus of child-directed English should yield sufficient data for the main ingredients of the native language of four year old children. Furthermore, it is plausible to conjecture that language acquisition never uses N's beyond a certain value, probably just a few hundred. It is doubtful that anyone can keep track of large values of N and e; perhaps learners will simply freeze the rule once they have seen enough data, i.e., a sufficiently large value of N. This implies that (a) the window of language acquisition will necessarily close at some point, and (b) it would be a mistake to use all the words we can get our hands on for productivity considerations (that is, no one learns morphology from the OED).

Third, the calibration of productivity under the TP deals with the type frequency of words, and more specifically, the proportion of e relative to N. It is immaterial exactly what these words are, or how frequently they appear — assuming, of course, they are frequent enough to be learned by young children.

Therefore, while different speakers will necessarily know different words, for obvious and non-obvious reasons, their grammar may still be the same if the relative proportions of N and e fall on the same side of productivity. This property brings about methodological convenience. In the absence of large child-directed speech corpora, we can word samples from a suitably large adult language corpus and calculate the productivity prediction for each sample. For rules whose developmental and productivity status is clear, we expect them to be consistently supported by the TP calculation for each sample. For instance, in the aforementioned child-directed English corpus (5 million words in total), there are 1,022 verbs that appeared in past tense, of which 127 are irregular (Yang 2016, 84). In comparison, of the top 1,022 verbs tagged as simple past in the British National Corpus (almost 200 million words), 128 are irregular. In the best case, productivity, as the balance between rules and exceptions and expressed by N and e, may be scale invariant across corpora as long as one stays in the relatively high frequency range.

Taken together, these facts of child language and the statistics of word distributions can provide useful tools for understanding language development from a quantitative perspective, even in the absence of perfect data. Consider a detailed study of the acquisition of the English dative constructions (Yang 2016, Section 6.3). The five-million-word corpus of child-directed English provides a reasonable coverage of the vocabulary for young children but the interesting cases, such as the well-known contrast between the ungrammatical "donate X Y" but the grammatical "assign X Y", involve vocabulary items learned much later than the stage represented in CHILDES. In fact, if the learner's experience is limited to the child-directed corpus — with just over 40 caused-possession verbs most of which are highly frequent (and monosyllabic) — the double object construction would be productive for this semantic class, thereby accounting for early developmental errors such as "I said them no", "I delivered you pizzas", "Should I whisper you something?". To get a fuller picture of the English speaker's knowledge of the construction, and to address the crucial problem of retreating from over-generalizations without negative evidence, we need to go beyond CHILDES and approximate the linguistic input of typical English speakers. Bootstrapping off CHILDES into the SUBTLEX corpus (Brysbaert and New 2009), I constructed a list of dative verbs, sorted by frequency, that would be relevant for the development of the double object construction.

Table 2: Caused-possession verbs and their expected distribution in the double-object construction

Top N	Yes	No	θ_N	Productive?
10	9	1	4	Yes
20	17	3	7	Yes
30	26	4	9	Yes
40	30	10	11	Yes
50	34	16	13	No
60	39	21	15	No
70	43	27	16	No
80	46	24	18	No
92	50	42	20	No

Table 2 shows that if a learner only learns from the most frequent dative verbs, such as those found in CHILDES, the double object construction will be deemed productive because the vast majority of these verbs — sufficiently many as assessed by the TP — will be attested in the construction. This corresponds to the early stage of over-generalization in child language. However, as the learner's vocabulary expands, as is approximated by including verbs with lower frequencies, the construction will no longer meet the productive threshold. The learner will thus retreat from the over-generalization and lexically memorize those verbs that do participate in the construction, although subdividing the verbs into finer semantic classes (e.g., "ballistic

motion" and "telecommunication" verbs) and applying the TP recursively may produce productivity generalizations (Yang 2016). The effective use of frequency/rank cutoff points, then, can provide a simulation of the developmental trajectory of rule acquisition.

3 Rules and Representations

With the advent of the Minimalist Program (Chomsky 1995) and the recent move away from the traditional highly complex conception of UG to domain-general principles of learning and computation (Yang 2002, Chomsky 2005, Berwick and Chomsky 2016, Yang et al. 2017), we have come full circle to a Piercian abductive learning framework that Chomsky alluded to in *Aspects* (1965) and *Language and Mind*: "The child cannot know at birth which language he is to learn, but he must know that its grammar must be of a predetermined form that excludes many imaginable languages. Having selected a permissible hypothesis, he can use inductive evidence for corrective action, confirming or disconfirming his choice. Once the hypothesis is sufficiently well confirmed, the child knows the language defined by this hypothesis; consequently, his knowledge extends enormously beyond his experience" (Chomsky 1968, p80).

Regarding the TP as the evaluation procedure that determines if a hypothesis is "sufficiently well confirmed", we naturally would like to know more about the hypothesis space and how it "excludes many imaginable languages". Methodologically, this is absolutely necessary to operationalize the TP: without knowing the format of the rule, we cannot count the items that could follow it or those that constitute exceptions.

A complete picture of the hypothesis space will likely elude us in the near future. But that should not be an impediment to making progress in empirical research which, in fact, should be accelerated by precise learnability principles such as the TP. On the one hand, the TP encourages the researcher to state precise generalizations about languages and language development, especially with respect to the division of labor between linguistic and non-linguistic constraints. On the other, it provides checks and balances to linguistic theorizing via its ability to detect significant linguistic generalization — assuming, of course, the essential correctness of the TP, a point to which I return in Section 4.

In practice, the researcher should approach an acquisition problem exactly the way they approach the problem as a linguist: every linguist was once a child. Given a set of data, what plausible generalizations would one draw? The analyst's task is aided by their knowledge of theoretical and comparative analysis of languages: crazy rules — skip every third phoneme, inverting the first auxiliary, etc. — presumably are crazy for children and linguists alike.

While principles such as structure dependence are plausibly universal, interesting questions arise when rules diverge in language specific ways. For instance, noun gender assignment can be semantically, phonologically, or morphologically conditioned, and there are languages where gender assignment is arbitrary not to mention languages without the use of gender at all. I will briefly summarize the case study of German noun plurals (Yang 2016, section 4.4) to illustrate the operationalization of the TP in acquisition studies.

The formulation of the TP suggests that a rule cannot be productive unless it has relatively few exceptions. On the face of it, this is inconsistent with the fact that sometimes minority rules can still be productive. For instance, the much-studied problem of German noun plural formation concerns the status of five suffixes (*-s*, *-(e)n*, *-e*, *-er*, and *-∅*). Notably, the *-s* suffix covers only a small minority of nouns. Table 3 provides the statistics based on some 450 highly frequent noun plurals in a corpus of child-directed German. The distribution is broadly similar to those based on larger corpora (e.g., Clahsen 1999). This is a significant fact in light of the discussion of word counting in Section 2: with some work, child-directed data, which is harder to come by, can be approximated with other corpora because the TP operates with type frequencies and relative proportions.

An application of the TP clearly will not identify a productive suffix in Table 3 since none is anywhere near requisite threshold. Yang (2016) reports several detailed case studies of this type: when children fail

Table 3: Distribution of noun plural suffixes for highly frequent nouns in child-directed German

suffix	types	%	% (Clahsen 1999)
-ø	87	18.9%	17%
-e	156	34.1%	27%
-er	30	6.5%	4%
-(e)n	172	37.5%	48%
-s	13	2.8%	4%

to discover a productive rule over a set of lexical items, they will subdivide the set along some suitable dimension and apply the TP recursively. For the German plural system, the relevant dimensions are gender and the phonological properties of the final syllable, which have been discussed in the previous literature on German morphology and phonology (e.g., Wiese 1996). Indeed, applying to the TP to the subdivided classes of nouns in Table 3 produces the correct results. The net effect is that almost all nouns are predictably accounted for by the four suffixes: each suffix will still have exceptions but the number of them fall under their respective tolerance threshold. This removes almost all nouns from consideration when it comes to the *-s* suffix, which has no structural restrictions on the noun and thus becomes the default.

In an acquisition study, however, the decision to partition nouns for recursive TP applications must be justified independently, and developmentally. That is, by the time young children show knowledge of the suffixes identified by the TP, they must have mastered the requisites that enable the partitions. Fortunately, in the study of German plurals, previous research had already established children's very early knowledge of noun genders. For instance, Mills (1986)'s classic study shows that German children across all age groups rarely produce gender-marking errors. This is similar to the other cases of morphological acquisition that most gender errors by children are those of omission rather than substitution. A more detailed study by Szagun (2004, 15) finds that the rate of correct gender marking is approximately 80% even before the age three, and it rises to nearly 100% before the age five. Interestingly, the 1986 study by Mills also notes that the acquisition of gender marking precedes, rather than follows, that of plural formation. This suggests that partitioning nouns according to gender is a logical as well as a developmental prerequisite for learning plural formation. This is a case of the TP forcing developmental questions and investigations that pertain to its application.

It should be pointed out that while linguists generally have good sense of what constitutes a plausible rule, they may be fallible when it comes to productivity. Statements about tendencies, which can made informally or even with quantitative measures, do not necessarily reflect the cognitive reality of rules. For instance, productivity has traditionally been associated with majority (e.g.,"statistical predominance"; Nida 1949, 45), and is still at the heart of virtually all probabilistic models of language learning. If that were correct, English would be a quantity-insensitive stress language as some 85% of words receive stress on the initial syllable (Cutler and Carter 1987), and no inflectional gaps (Halle 1973, Baerman et al. 2010) would ever emerge because surely one of the allomorphs is the most frequent and thus statistically dominant. To use a familiar example, the prefix *un-* has been used as a diagnostic to identify verbal vs. adjectival passives (Wasow 1977, Levin and Rappaport 1986, Embick 2004) but a careful look at the data reveals that *un-* is in fact unproductive and must lexically select the stem it attaches to (Yang 2018). Hence we have *missed/*unmissed opportunities*, *recommended/*unrecommended hotels*, *split/*unsplit bills*, etc., and no one says *unblack*, *unquick*, or *ungreat*.

This last point is important because theoretical frameworks often diverge *because of* productivity: see Lakoff (1965) and Chomsky (1970), and more recently Marantz (1997) and Williams (2007). Depending

on the theory, productive and unproductive processes may receive different representations and reside in different components of the grammar. The TP provides an alternative to the purely structural analysis more familiar in theoretical linguistics. Presumably both the child and the linguist need to identify significant generalizations about language: what the TP offers is a benchmark for what counts as significant, the chief among which is the infinite productivity of language. Moreover, the TP operationalizes productivity research at the level of plausible generalizations about the data without making unnecessary theoretical commitments: "add -ed to verbs to form past tense" can be stated either "in the lexicon" via word formation rules (Anderson 1992) or "in the syntax" (Halle and Marantz 1993), and the bean counting of N, e, and θ_N is all the same. At the minimum, the TP provides a lower bound on what is distributionally learnable from data. If this approach is successful, then explanatory adequacy no longer resides in the intricacies of theory-internal apparatus or principles and constraints specific to language.

4 Unreasonable Effectiveness

So far I have taken the TP as axiomatic and discussed its implication for language and language acquisition research. A confession and a disclaimer are now in order.

The TP has been unreasonably, and puzzlingly, effective. Its derivation relies on several idealized assumptions about language, including the Zipfian distribution of word frequencies which, while generally accurate when verified against large corpora, is never exactly true. Thus the effectiveness of the TP in the empirical studies of language acquisition is surprising. It has been applied successfully hundreds of time by myself and others, including many case studies reported in Yang (2016): not a single time would the idealized assumptions hold strictly true and generally no one even bothered checking. This point is best illustrated in artificial language learning experiments: words and their frequencies are under strict control and children's vocabulary and outcome of learning (i.e., the status of rules) can be assessed at the individual level.

For example, in Schuler et al. (2016)'s study, young children learn nine novel nouns. In one condition, five nouns share a plural suffix ("regulars"), and the other four are idiosyncratic ("irregulars"). In the other condition, the mixture is three regulars and six irregulars. The choices of 5R/4E and 3R/6E are by design: the TP predicts the productive extension of the regular suffix in the 5R/4E condition because four exceptions are below the threshold ($\theta_9 = 4.2$) but there is no generalization in the 3R/6E conditions. In the latter case, despite the statistical dominance of the regular suffix, the six exceptions exceed the threshold. When presented on additional novel items in a Wug-like test, almost all children in the 5R/4E condition generalized in a process akin to the productive use of English "-ed", and none in the 3R/6E condition did, much like speakers trapped in morphological gaps without a productive rule (Halle 1973, Gorman and Yang 2018).

These robust results are unexpected. The derivation of the TP uses a well-known approximation about H_N, the N-th Harmonic number, which appears in the Zipfian assumption of word frequencies:

$$H_N \approx \ln N$$

The approximation works well only when N is very large. For small values of N, the discrepancies are considerable. For $N = 9$, as in Schuler et al. (2016), the exact calculation of H_N produces the threshold of 3 whereas the (crude) approximation of $\ln N$ produces the threshold of 4 — which is in fact what children categorically use. Furthermore, while the experiments tried to approximate the Zipfian distribution of words, the match is not exact: Zipf's Law has a characteristic long tail of words that appear only once but a word appearing only once in an artificial language is very difficult for young children to learn. Even more strikingly, if children were to use actual word frequencies to calibrate the productivity of rules, they would *not* have discovered the productive rule. In one of the 5R/4E conditions, the five regulars are the five most frequent words. It is easy to see that the more efficient grammar is actually not to have a productive rule: all

nine nouns ought to be listed. Having a productive rule would force the five most frequent regulars to "wait" for the four least frequent irregulars, which is in fact slower than having full listing and no productive rule at all. Yet children chose the productive rule option categorically.

Another strong confirmation can be found in Emond and Shi (2021). These authors designed two sets of stimuli, each of which consisted of 16 distinct items. In the first set, 11 out of the 16 items followed a word order pattern; in the second, 10 out of the 16 items followed the pattern. The design reflects the prediction of the TSP. For N=16 items, the critical threshold is $\theta_{16} = 5.77$: 10 is insufficient for generalization despite being the majority but 11 is. Indeed, 14-month-old infants generalized the pattern from the 11/16 set, but not the 10/16 set. In this study, as in a previous experiment (Koulaguina and Shi 2019), the words were introduced to children with uniform, rather than Zifpian, frequencies, thereby violating a fundamental assumption underlying the derivation of the TP. Yet children's behavior is again well predicted.

At the moment it is not clear why the TP should work as well as it does: an idealized model of the reality does better than reality itself. Nor is clear how the brain actually implements something like the TP. It seems likely that children only keep track of the relative ratio of e and N, but cannot keep track of these exact values. I is virtually certain that these quantities cannot be precisely represented, at least not explicitly: it is extremely unlikely if children could report back how many rule-following and rule-defying items they have learned during a brief experiment. Yet somehow children make a category decision on the basis of the relative magnitude of two quantities.

A threshold such as the TP may not be empirically identifiable from observations. Suppose one collects many examples of productive and unproductive rules in many languages in the hope of discovering what the threshold for productivity may be. For each rule, however, the actual number of exceptions may be arbitrarily distant from the productivity threshold: the precise value cannot be reliably determined.

Meanwhile, the TP offers protections against over-fitting: As long as the number of exceptions is below the threshold, the learner need not improve on their coverage of the data. This ensures that individual learners in similar linguistic environments will essentially learn the same grammar, even though their individual vocabulary and experiences will be significantly different.

References

Anderson, S. R. (1969). *West Scandinavian vowel systems and the ordering of phonological rules*. PhD thesis, MIT.

Anderson, S. R. (1992). *A-morphous morphology*. Cambridge University Press, Cambridge.

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

Baerman, M., Corbett, G. G., and Brown, D., editors (2010). *Defective paradigms: Missing forms and what they tell us*. Oxford University Press, Oxford.

Berwick, R. C. and Chomsky, N. (2016). *Why only us: Language and evolution*. MIT Press, Cambridge, MA.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

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

Carlson, M. T., Sonderegger, M., and Bane, M. (2014). How children explore the phonological network in child-directed speech: A survival analysis of children's first word productions. *Journal of memory and language*, 75:159–180.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Chomsky, N. (1968). *Language and mind*. Harcourt, Brace and World.

Chomsky, N. (1970). Remarks on nominalization. In Jacobs, R. A. and Rosenbaum, P., editors, *Readings in English transformational grammar*, pages 184–221. Ginn, Waltham, MA.

Chomsky, N. (1995). *The minimalist program*. MIT Press, Cambridge, MA.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, 36(1):1–22.

Clahsen, H. (1999). Lexical entries and rules of languge: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22:991–1069.

Cutler, A. and Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2(3–4):133–142.

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

Embick, D. (2004). On the structure of resultative participles in english. *Linguistic Inquiry*, 35(3):355–392.

Emond, E. and Shi, R. (2021). Infants' rule generalization is governed by the Tolerance Principle. In Dionne, D. and Vidal Covas, L.-A., editors, *Proceedings of the 45nd annual Boston University Conference on Language Development*, pages 191–204.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., and Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, pages i–185.

Fenson, L., Marchman, V., Thal, D. J., Dale, P. S., Reznick, J. S., and Bates, E. (1993). *The MacArthur communicative development inventories. User's guide and technical manual*. Singular, San Diego.

Goodman, J. C., Dale, P. S., and Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of child language*, 35(3):515–531.

Gorman, K. and Yang, C. (2018). When nobody wins. In Rainer, F., Gardani, F., Luschützky, H. C., and Dressler, W. U., editors, *Competition in inflection and word formation*, pages 169–193. Springer, Berlin.

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

Halle, M. and Marantz, A. (1993). Distributed morphology and the pieces of inflection. In Hale, K. and Keyser, S. J., editors, *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger*, pages 111–176. MIT Press, Cambridge, MA.

Hart, B. and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, Baltimore, MD.

Kiparsky, P. (1973). Elsewhere in phonology. In Anderson, S. R. and Kiparsky, P., editors, *A festschrift for Morris Halle*, pages 93–106. Holt, Rinehart and Winston, New York.

Koulaguina, E. and Shi, R. (2019). Rule generalization from inconsistent input in early infancy. *Language Acquisition*, 26(4):416–435.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia.

Labov, W., Ash, S., and Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology, and sound change*. Mouton de Gruyter, Berlin.

Lakoff, G. (1965). *On the nature of syntactic irregularity*. PhD thesis, Indiana University.

Levin, B. and Rappaport, M. (1986). The formation of adjectival passives. *Linguistic inquiry*, pages 623–661.

Lignos, C. and Yang, C. (2016). Morphology and language acquisition. In Hippisley, Andrew R. abd Stump, G., editor, *The Cambridge handbook of Morphology*, chapter 28, pages 765–791. Cambridge University Press, Cambridge.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum, Mahwah, NJ, 3rd edition.

Marantz, A. (1997). No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. In Dimitriadis, A., Siegel, L., Surek-Clark, C., and Williams, A., editors, *Penn Working Papers in Linguistics 4.2: Proceedings of the 21st annual Penn Linguistics Colloquium*, pages 201–225. Penn Linguistics Club, Philadelphia.

Marcus, G., Pinker, S., Ullman, M. T., Hollander, M., Rosen, J., and Xu, F. (1992). *Overregularization in language acquisition*. Monographs of the Society for Research in Child Development. University of Chicago Press, Chicago.

Mills, A. (1986). *The acquisition of gender: A study of English and German*. Springer, Berlin.

Nagy, W. E. and Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, 19(3):304–330.

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

Nida, E. A. (1949). *Morphology: the descriptive analysis of words*. University of Michigan Press, Ann Arbor, 2nd edition.

Plunkett, K. and Marchman, V. A. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48(1):21–69.

Rowe, M. L. and Goldin-Meadow, S. (2009). Differences in early gesture explain ses disparities in child vocabulary size at school entry. *Science*, 323(5916):951–953.

Rowe, M. L., Raudenbush, S. W., and Goldin-Meadow, S. (2012). The pace of vocabulary growth helps predict later vocabulary skill. *Child development*, 83(2):508–525.

Schuler, K. (2017). *The acquisition of productive rules in child and adult language learners*. PhD thesis, Georgetown University, Washington, D.C.

Schuler, K., Yang, C., and Newport, E. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. In *The 38th Cognitive Society Annual Meeting*, Philadelphia, PA.

Szagun, G. (2004). Learning by ear: on the acquisition of case and gender marking by German-speaking children with normal hearing and with cochlear implants. *Journal of Child Language*, 31(1):1–30.

Wasow, T. (1977). Transformations and the lexicon. In Culicover, P. W., Wasow, T., and Akmajian, A., editors, *Formal Syntax*, pages 327–360. Academic Press, New York.

Wiese, R. (1996). *The phonology of German*. Clarendon, Oxford.

Williams, E. (2007). Dumping lexicalism. In Ramchand, G. and Reiss, C., editors, *The Oxford handbook of linguistic interfaces*, pages 353–382. Oxford University Press, Oxford.

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

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

Yang, C. (2018). The said and the unsaid: A modern look at English verbal and adjectical passives. In Hollebrandse, B., Kim, J., Pérez-Leroux, A. T., and Schulz, P., editors, *T.O.M and Grammar: Thoughts on Mind and Grammar: A Festschrift in Honor of Tom Roeper* , pages 195–199. University of Massachusetts, Amherst, MA.

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., and Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews*, 81(Part B):103 – 119.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA.

REPLY

Some consequences of the Tolerance Principle

Charles Yang

The commitment to a formalist theory of language acquisition seems to have resonated with the commentators. In what follows, I will discuss and expand on some of the central issues surrounding the Tolerance Principle (TP).

1. Clarifications and corrections

The book-length treatment of the TP (Yang, 2016) would have provided a fuller background and assuaged some of the commentators' concerns. For instance, both **Wittenberg and Jackendoff** and **Kapatsinski** are incredulous that language could operate like a serial search model, the algorithmic foundation of the TP – because, they argue, the brain is parallel. But is a parallel brain really incompatible with serial behavioral effects? My (presumably parallel) brain can memorize and recite the digits of π in a strictly linear sequence. And there are numerous serial effects in the study of cognition that are the product of a parallel brain: the scanning memory model of Sternberg (1969), the linear search model of number representation and processing (Gallistel et al., 1992; Brannon et al., 2001), not to mention Weber's Law (Gibbon, 1977). More to the point, a serial model is simply a better account of lexical processing. A picture (actually two) is worth a thousand words.

The results are based on lexical decision data from almost 40,000 words (Balota et al., 2007). Factors affecting reaction time (word length, orthographic neighborhood size, etc.) have been controlled ("residualized"; see Lignos, 2013 for details). Word rank (bottom) clearly provides a better fit than word frequency (top). Incidentally, even frequency-based accounts always use the logarithm of frequency, which brings it surreptitiously close to rank.

I do agree with Wittenberg and Jackendoff that one should always pursue multi-level explanations (in the sense of Marr). I have done so myself (Yang, 2016, 76–78, Yang, 2017) while criticizing Bayesian rational analysis models that explicitly disavow psychological reality – which, confusingly, Kapatsinski

https://doi.org/10.1075/lab.00022.yan
Linguistic Approaches to Bilingualism 8:6 (2018), **797–809**.
ISSN 1879–9264 | E-ISSN 1879–9272

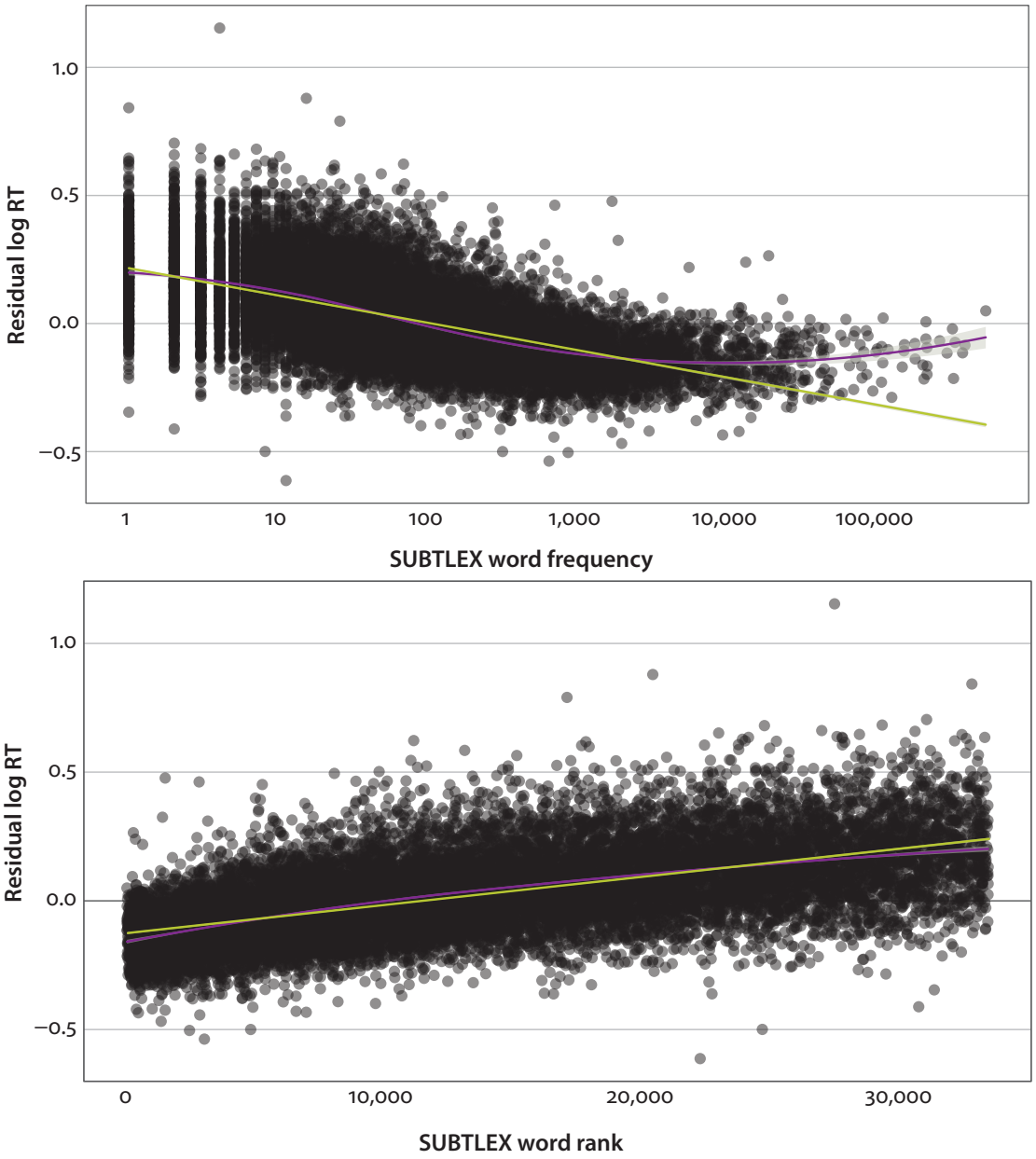

Figure 1. Comparison of word frequency and rank as predictor of lexical decision time. From Lignos (2013)

embraces as an article of faith despite his concern for psychological grounding. The appropriate response should be to develop a neural theory of serial effects, rather than disregarding serial effects from behavioral studies just because 'the brain is parallel', especially when these commentators don't even offer insight on how the parallel brain implements *parallel* effects.

Kapatsinski seems to have read the book but only selectively. He completely ignored my thorough cross-linguistic review of morphological acquisition, which shows unambiguously that productivity is a categorical effect (Xu & Pinker, 1995).

Rather, he prefers his own study of adult artificial language learning where the subjects, on average, produced a gradient score on a Wug test and concludes that productivity must be gradient. But we already know that adults do not approach (artificial) language learning tasks in the categorical way that children do (Hudson Kam & Newport, 2005); more in Section 2. And it's important to recall that the original Wug study (Berko, 1958) already demonstrated the task-specific differences between children and adults. The subjects were presented with novel verbs such as spow and bing, which are similar to existing irregular verbs (blow-blew, sing-sang). Only one of the 86 children tested produced bang (no said said spew, and only spowed was produced). Adults, however, are far more willing to produce irregularized forms, which is likely to a task effect, since there have been virtually no such cases in the natural history of English verbs (Anderwald, 2013; Yang, 2016).

Goldberg also read the book but doesn't seem to have understood it. She wonders how one learns the prefix *pre-* as in *pre-Watergate, pre-Trump,* etc. According to her, since children will learn many nouns and proper names but presumably relatively few will appear with *pre-,* the productivity threshold would not be reached and the prefix cannot be learned. This is a perverse reading of the TP and the general problem of inference. Just learning a word does not force the learner to evaluate all conceivable ways in which the word is be used. The past tense *-ed* can only be learned when (enough) verbs have appeared in past tense; the suffix wouldn't even be entertained when the child hears and uses a verb in the present. In the acquisition study of the dative constructions summarized in the target article, ditransitive verbs such as *slip* are not included because every instance of it in the child-directed input corpus is an intransitive form (e.g., *They slipped).* By the same token, not every noun or proper name – thank goodness – would reach the notoriety of *Watergate* and *Trump,* so *truck, milk,* and *Abby* do not factor into the calibration of *pre-* at all. Goldberg seems to believe that if a form (e.g., "pre-tortilla") *can* be used, it *must* be used. This clearly doesn't follow but it does explain her persistent appeal to indirect negative evidence (e.g., Boyd & Goldberg, 2011), the contrapositive of the above: If a form is *not* used, it must *not* be grammatical – which fails empirically as well (Yang, 2015). In the rest of her commentary, Goldberg summarizes her own proposal: "(P)reviously witnessed partially-abstracted exemplars cluster together in our hyper-dimensional representational space for language, forming a massively interrelated dynamic system (a construct-i-con), which is simply an expanded version of what has long been recognized to be needed for our knowledge of words (the lexicon)" (p. 729). I have no idea what this means. While I'm quite prepared for someone to show the TP to be wrong – so long as as they know how to use it right – it should really be replaced with a better equation, not vague analogies and metaphoric allusions.

I am pleased that two prominent usage-based researchers, **Gries** and **Rowland,** agree with my call for methodological rigor: Gries has made similar pleas and Rowland even gives me a Popperian endorsement. The TP, so far as I can tell, is an example of the learning mechanism that has been viewed as the central goal of usage-based researchers: "a single mechanism responsible both for generalization, and for restricting these generalizations to items with particular semantic, pragmatic, phonological (and no doubt other) properties (Ambridge & Lieven, 2011, p. 267)." A useful common ground.

But their defense of the usage-based position is unconvincing. **Gries** proposed a log odds ratio measure which shows the frequency of *give me* is higher than "expected by chance", and is thus "at least compatible with the notion that *gimme* might be a unit" (p. 735). But neither point is correct. It is true that *me* follows *give* more frequently than "by chance", with "chance" understood as "other words" (Gries; Table 1). But the only way to establish productivity is statistical independence; i.e., the frequency of *give me* can be predicted from the frequency of *give* and that of *me.* The "other words" do not matter. Furthermore, what if give me is indeed abnormally frequent? It still doesn't follow that give me is a holistic unit. Frequency and compositionality are in principle independent of each other. Gries seems to uphold the idea that whole-unit frequency effects – if real, though not in the present case – automatically counts against compositionality. This is a dogma from the past tense debate as pointed out in the target article; see Yang (2002), Taft (2004), Fruchter and Marantz (2015), Regel et al. (2015) for acquisition, processing, and neurological evidence for compositional approaches to whole-unit frequency effects.

Rowland does not directly challenge the statistical findings of my determiner work (Yang, 2013a) but brings up the study of Pine et al. (2013). In some subsamples of child language, children are assessed to have a lower overlap score than adults. Rowland faults me for not discussing this result; here is why. The Pine et al. (2013) method is biased, and generally undervalues the productivity of the smaller sample, which is usually the child corpus because adults talk more. One can develop an analytical diagnosis of the bias – too complex to summarize here (see Yang & Valian, 2018) – but the problem can also be revealed by a minimum sanity check, on samples produced by (adult) speakers whose knowledge of the determiners is not in question. Doing so would have shown Mary MacWhinney to be (absurdly) less productive than Brian (the curator of CHILDES) in the MacWhinney corpus: Brian just talked a lot more. Once again, quantitative results are interpreted at face value, and methods that produce (preferred) results are deployed without validation.

2. Children, adults, and vocabularies

Is there any continuity between L1 and L2, not to mention those cases – "atypical" development, simultaneous and sequential early multiple acquisition, heritage language and attrition, etc. – that lie in between? This is obviously too large of question and my goal is much more specific. I propose that the TP is operative for adult language learning, and its smaller-is-better property, that rules are easier than learn when the relevant vocabulary is smaller, can account for adult's evident deficiency in comparison to children.

How do we show that the TP is used by adults at all? There is *prime facie* evidence that suggests otherwise. For example, in a series of studies (summarized in Schuler, 2017), subjects learn artificial languages where rules have various levels of exceptions. Children follow the TP nearly categorically but adults generally match the token frequencies of the available forms in the stimuli.

It remains unclear why adults probability match while children prefer categorical rules, a difference found in other domains of learning and decision making (e.g., Weir, 1964; Derks & Paclisanu, 1967). Yet there are at least two reasons to believe that the TP holds for learners of all ages. Theoretically, the central assumption of the Tolerance Principle, namely the Elsewhere Condition, is not known to degrade across development. Empirically, there is evidence that adults, and late child learners, can learn rules extremely well. First, a significant portion of English derivational morphology is acquired quite late (Tyler & Nagy, 1989; Jarmulowicz, 2002), presumably because it involves advanced vocabulary that comes with literacy and education. Second, L2 learners *can* form productive rules in a manner similar to L1 learners (White & Genesee, 1996). Yang and Montrul (2017) provide an extensive review of L2 acquisition of the English dative constructions. These constructions are informative because their grammatical properties are obscure and most English teachers, I'd imagine, would never offer lessons on the distribution of *donate*. But L2 learners also go through the stages of over-generalization and retreat like L1 learners, and they gradually refine the phonological and semantic restrictions on the constructions over time, with advanced learners showing native-like grammaticality judgment (Jäschke & Plag, 2016)

The commentators are correct to stress the complexity of L2 acquisition. **Paradis** highlights the individual differences in L2 that cannot be attributed to "language-level" factors such as word frequencies. **Dominíguez and González Alonso, Montrul,** and **Yusa** point out that the input for L2 acquisition is filtered through the learner's L1. The target article recognizes these complications. For instance, I chose adult Italian learners of English to demonstrate the presence of a topic-drop grammar (á la Chinese and Japanese; see also Yusa, this volume), which is neither in the speaker's L1 (pro-drop) or L2 (obligatory subject), thereby

providing unambiguous evidence for Full Access. Similarly, the analysis of determiner productivity in L2 shows comparable statistical results for Italian and Punjabi speakers despite the differences in their L1 determiner systems, which should address **Dimroth**'s concerns. And the TP, with its focus on vocabulary composition, is well equipped to handle both language- and individual-level differences. The relative ease of French past tense acquisition noted by Paradis would follow my account of why the English plural suffix *-s* is learned earlier than the past tense *-ed* (fewer exceptions; Yang, 2016, 4.1.3). And the onset of rule productivity for English-learning children Adam, Eve, and Abe can be predicted by their vocabulary acquisition (Yang, 2016, 4.1.2).

Under the TP, the effect of L1 on L2 is formally no different from (purely) L1 acquisition. It is trivially true that the child doesn't learn everything they hear; otherwise they would learn 50,000 words by the age of two. But just saying the input is filtered (**Biberauer, Perkins & Lidz**) does not solve anything; one still needs to understand how rules come out of the "intake". I think there is little prospect in a general theory of filtering because the input may be reduced by a virtually unlimited range of factors: cognitive limitations in children, L1 influence in L2 adults, a kid overly obsessed with Lego, an ESL student who Facebooks rather than paying attention in class. But the learner's vocabulary composition, both L1 and L2, can be estimated and the TP will makes clear claims about grammatical rules no matter how filtering works.

On the matter of vocabulary, several commentators (**De Cat, Dimroth, Slabakova**) question my take on less-is-more. I should have been more clear: while a smaller vocabulary does make rule learning easier, it still needs to be large enough for the rule to be learnable (e.g., enough regular verbs to overcome the irregulars). Thus a younger learner may not be better than an older one at everything: I have already discussed the gradual refinement of the English dative constructions because the requisite vocabulary can only be built up over time, so older learners are "better" than younger ones. Thus, **Dimroth**'s interesting study that child L2 learners perform better than L1 learners on German verbal morphology is perfectly compatible with the TP: although a fine grained corpus analysis is necessary, the complexity of the German system would seem to require a substantial vocabulary which an older child may acquire faster. And it is definitely *not* the case that bilingual children would learn rules faster than monolingual learners, contrary to **De Cat**'s understanding: reduced input as in the case of bilingual acquisition will reduce the vocabulary necessary to support rule productivity, which is exactly what Marchman et al. (2010) find.

A related question, raised by **De Cat, Dimroth,** and **Kapatsinski** in somewhat different ways, concerns the completion of rule learning. Since the TP requires a great majority of words to follow a rule to ensure productivity, waiting too long

before coming to a decision (i.e., with a large N) would render every rule unproductive because of the data sparsity (Yang, 2013b). The answer comes in two ways, both suggesting that the learner will stop looking and "freeze" the rules in place at a value of *N* no more than a few hundred. Empirically, a three-year-old's vocabulary size is no more than just around 1,000 (Hart & Risley, 1995). This is at an age where the core grammar (word order, inflectional morphology, etc. though not everything) is already solidly in place. Thus, productivity decisions can, and must, be made when *N* is quite modest. It is important to stress that the learning limit of *N* is not a function of age: the full details of the dative constructions are learned quite late but the value of *N* for the verbs is probably no more than 100; see also the discussion of L2 datives above. Conceptually, as I discussed elsewhere (Yang, 2016, 76ff, Yang, 2018), the TP has been surprisingly, and unreasonably, effective, especially because the numerical assumptions in its derivation are almost never strictly true (and no one even bothers checking). It seems that children somehow keep track of two quantities and compute their relations. It is difficult to envision high-precision calculation for large *N*'s although the neural implementation of something like the TP is completely unknown.

3. Learnability and the theory of grammar

The last set of comments comes from theoretical linguists or acquisition researchers who make a strong ontological commitment to theoretical linguistics. Some worry whether the TP has gone too far in the other direction, without paying sufficient attention to the representational and other constraints in the grammar (**De Cat, Dominíguez & González Alonso, Perkins & Lidz, Roeper, Slabakova).**

My general approach is to have as little UG as possible (Berwick & Chomsky, 2016). The application of the TP has been, by design, based on what can be described as plausible generalizations about the data without making (unnecessary) theoretical commitments about how such generalizations are to be stated. For instance, "add *-ed* to verbs to form past tense" can be stated either "in the lexicon" or "in the syntax" – a matter of fierce theoretical controversy but the bean counting of *N*, *e*, and θ_N is all the same. The TP provides a lower bound on what is distributionally learnable from data. If this approach is successful, then explanatory adequacy no longer resides in the intricacies of theory-internal apparatus or principles and constraints specific to language (Yang et al., 2017).

This will invite skepticism. **Perkins and Lidz** believe my theory fails to take developmental constraints into account. They also question children's ability to detect semantic properties, e.g., caused-possession in the dative constructions. For them, it is the syntax that helps the learner to narrow down the semantics

(syntactic bootstrapping; Gleitman, 1990). But they don't seem to realize that the TP already provides a *developmental* theory of syntactic bootstrapping: syntax does help with semantics but syntax has to be learned.[1] Table 3 of the target article shows how the vocabulary of dative verbs, and thus the double-object construction, grows over time. Let's grant that children can't "observe" the meaning of verbs such as *promise* and *guarantee* without syntax (although Perkins & Lidz only offer assertions to this effect without evidence). The syntax for bootstrapping can be formed when the vocabulary is small and contains only "easy" words (Gleitman et al., 2005) – *give, feed, hand, show, bring* – whose meanings *are* observationally learnable (Trueswell et al., 2016). This provides the TP-sanctioned inductive basis that caused possession is encoded in the double-object structure, with which children can decode the meanings of *promise* and *guarantee*.

Perkins and Lidz are also concerned that my approach may "miss important generalizations about language acquisition" (p. 743) such as "(I)f a language has two ditransitive constructions, the one expressing caused possession is always the one in which the goal c-commands the theme … And, children seem to know this link despite a severe poverty of evidence" (p. 746). I fail to see the relevance. How does knowing the goal c-commanding the theme help learn that *donate* cannot appear in the double-object construction but *assign* can? Never mind the supposed generalization is false: Middle English (Visser, 1963) and modern Scandinavian languages to varying degrees, allow both the goal-theme and theme-goal word order. The same holds for the suggestion that rule learning may be aided by features and other formal structures (**Biberauer, Dominínez & González Alonso, Slabakova).** Perhaps productive and unproductive processes are indeed differentiated representationally but that is clearly the result of learning not a theory of learning, e.g., which words belong where on the hierarchy, which features become general and "abstract" and which are conservative and lexically specific. In this vein, **Svenonius** raised the problem of object shift in Norwegian, where children fail to consistently shift in obligatory contexts. The distribution of shifted objects can be described in different theoretical frameworks with some more surface-oriented that others but the learning problem is the same and has already been subjected to a TP analysis (Anderssen et al., 2012, 57). The number of shifted object pronominals is only a small subset of all (10/39): not shifting is "productive" and children must memorize those that do. Failing to shift consistently is expected because exceptions may be regularized as in the familiar case of English past tense. Similarly, the obligatory use of determiners in languages like Italian needn't follow from the property of some null head – and one would need a story

1. I thank Lila Gleitman for discussions of this matter over some funky blue bread pudding.

of why it is *not* available for English – but can be learned distributionally from input (Ceolin, 2018).

But Svenonius's general message is important: what are the "constraints on the 'format' of lexical items therefore define the hypothesis space" (p. 781). UG won't provide a complete set of the primitives to structure the hypothesis space. It is inconceivable that the noun classes in Bantu, the classifier system in Japanese, and the past tense rules for the irregular verbs in English are all carved out of the innate universal template. More likely, these linguistic categories are established because children can discover, using something like the TP, formal correspondences that relate them. A case in point is the "telecommunication" subclass of dative verbs, which is surely not an innate semantic class but one established on their participation in a formal structure namely the double object construction. The child is probably innately primed to organize the categories in a combinatorial system ("features"), which may follow from Merge and perhaps other general principle of system organization (e.g., the particulate principle; Studdert-Kennedy, 1998).

4. Conclusion

The TP provides a new division of labor between what can be learned and what needs to be built in. As **Rothman and Chomsky** point out, this can eliminate "arbitrary stipulations of parameter values" (p. 765) and provides an account of the idiosyncractic properties of particular grammars without overburdening the biological requirement for language. Indeed, the minimalist approach (Berwick & Chomsky, 2016) encourages a return to an earlier, abductive, framework of language acquisition: "Having selected a permissible hypothesis, he can use inductive evidence for corrective action, confirming or disconfirming his choice. Once the hypothesis is sufficiently well confirmed, the child knows the language defined by this hypothesis; consequently, his knowledge extends enormously beyond his experience" (Chomsky, 1968, p. 80). The TP determines whether a hypothesis is "sufficiently well confirmed".

It seems appropriate to end with the concluding remarks from Rothman and Chomsky, who quote Chomsky (1995): "The field is changing rapidly under the impact of new empirical materials and theoretical ideas. What looks reasonable today is likely to take a different form tomorrow. … Whether these steps are on the right track or not, of course, only time will tell" (p. 9). This will take a collective endeavor from many theoretical and empirical angles as the commentators have helpfully made clear.

References

Ambridge, B. & Lieven, E. V. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511975073

Anderssen, M., Bentzen, K., & Rodina, Y. (2012). Topicality and complexity in the acquisition of norwegian object shift. *Language Acquisition*, 9(1), 39–72. https://doi.org/10.1080/10489223.2012.633844

Anderwald, L. (2013). Natural language change or prescriptive influence?: Throve, dove, pled, drug and snuck in 19th-century American English. *English World-Wide*, 34(2), 146–176.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13(3), 263–308. https://doi.org/10.1016/0010-0277(83)90012-4

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3), 445–459. https://doi.org/10.3758/BF03193014

Berko, J. (1958). The child's learning of English morphology. *Word*, 14(2–3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

Berwick, R. C. & Chomsky, N. (2016). *Why only us: Language and evolution*. Cambridge, MA: MIT Press.

Biberauer, T. (2018). Less is More: On the Tolerance Principle as a manifestation of Maximize Minimal Means. *Linguistic Approaches to Bilingualism*, 8(6), 707–711.

Boyd, J. K. & Goldberg, A. E. (2011). Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language*, 87(1), 55–83. https://doi.org/10.1353/lan.2011.0012

Brannon, E. M., Wusthoff, C. J., Gallistel, C., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science*, 12(3), 238–243. https://doi.org/10.1111/1467-9280.00342

Ceolin, A. (2018). Explaining cross-linguistic differences in article omission through an acquisition model. In Bertolini, A. B. & Kaplan, M. G., (Eds.), *Proceedings of the 42nd annual Boston University Conference on Language Development*, pp. 100–113, Somerville, MA: Cascadilla Press.

Chomsky, N. (1968). *Language and mind*. Harcourt, Brace and World.

Chomsky, N. (1995). *The minimalist program*. Boston: MIT Press.

Derks, P. L. & Paclisanu, M. I. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology*, 73(2), 278. https://doi.org/10.1037/h0024137

De Cat, C. (2018). Evaluating Yang's algorithms: An outline. *Linguistic Approaches to Bilingualism*, 8(6), 712–716.

Dimroth, C. (2018). Input and the acquisition of productive grammatical knowledge: Vocabulary size as missing link? *Linguistic Approaches to Bilingualism*, 8(6), 717–721.

Domínguez, L. & González Alonso, J. (2018). What is the role of L1 representations in a grammar-input model of L2 acquisition? *Linguistic Approaches to Bilingualism*, 8(6), 722–726.

Fruchter, J. & Marantz, A. (2015). Decomposition, lookup, and recombination: MEG evidence for the Full Decomposition model of complex visual word recognition. *Brain and Language*, 143, 81–96. https://doi.org/10.1016/j.bandl.2015.03.001

Gallistel, C. R. & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, 44(1–2), 43–74. https://doi.org/10.1016/0010-0277(92)90050-R

Gibbon, J. (1977). Scalar expectancy theory and weber's law in animal timing. *Psychological Review*, 84(3), 279. https://doi.org/10.1037/0033-295X.84.3.279

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1(1), 3–55. https://doi.org/10.1207/s15327817la0101_2

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, 1(1), 23–64. https://doi.org/10.1207/s15473341lld0101_4

Goldberg, A. (2018). The sufficiency principle hyperinflates the price of productivity. *Linguistic Approaches to Bilingualism*, 8(6), 727–732.

Gries, S. Th. (2018). Mechanistic formal approaches to language acquisition: Yes, but at the right level(s) of resolution. *Linguistic Approaches to Bilingualism*, 8(6), 733–737.

Hart, B. & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing, Baltimore, MD.

Hudson Kam, C. L. & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2), 151–195. https://doi.org/10.1080/15475441.2005.9684215

Jarmulowicz, L. (2002). English derivational suffix frequency and children's stress judgements. *Brain and Language*, 81(1–3), 192–204. https://doi.org/10.1006/brln.2001.2517

Jäschke, T. & Plag, I. (2016). The dative alternation in German-English interlanguage. *Studies in Second Language Acquisition*, 38, 485–521. https://doi.org/10.1017/S0272263115000261

Kapatsinski, V. (2018). On the intolerance of the Tolerance Principle. *Linguistic Approaches to Bilingualism*, 8(6), 738–742.

Lidz, J. & Perkins, L. (2018). The importance of input representations. *Linguistic Approaches to Bilingualism*, 8(6), 743–748.

Lignos, C. (2013). Modeling words in the mind. PhD thesis, University of Pennsylvania.

Marchman, V. A., Fernald, A., & Hurtado, N. (2010). How vocabulary size in two languages relates to efficiency in spoken word recognition by young spanish-english bilinguals. *Journal of Child Language*, 37(4), 817–840. https://doi.org/10.1017/S0305000909990055

Montrul, S. (2018). Learning a Second Language Takes More than Math. *Linguistic Approaches to Bilingualism*, 8(6), 749–752.

Paradis, J. (2018). Language-level input factors are not enough to explain child bilingual acquisition. *Linguistic Approaches to Bilingualism*, 8(6), 753–757.

Pine, J. M., Freudenthal, D., Krajewski, G., & Gobet, F. (2013). Do young children have adultlike syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3), 345–360. https://doi.org/10.1016/j.cognition.2013.02.006

Regel, S., Opitz, A., Müller, G., & Friederici, A. D. (2015). The past tense debate revisited: Electrophysiological evidence for subregularities of irregular verb inflection. *Journal of Cognitive Neuroscience*, 27(9), 1870–1885. https://doi.org/10.1162/jocn_a_00818

Roeper, T. (2018). Grammar acquisition and grammar choice in the variationist model. *Linguistic Approaches to Bilingualism*, 8(6), 758–763.

Rothman, J. & Chomsky, N. (2018). Towards eliminating arbitrary stipulations related to parameters: Linguistic innateness and the variational model. *Linguistic Approaches to Bilingualism*, 8(6), 764–769.

Rowland, C. (2018). The principles of scientific inquiry. *Linguistic Approaches to Bilingualism*, 8(6), 770–775.

Schuler, K. (2017). The acquisition of productive rules in child and adult language learners. PhD thesis, Georgetown University, Washington, D.C.

Schütze, C. T. (2005). Thinking about what we are asking speakers to do. In Kepser, S. & Reis, M., (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 457–485. Mouton de Gruyter, Berlin. https://doi.org/10.1515/9783110197549.457

Slabakova, R. (2018). Back to our roots. *Linguistic Approaches to Bilingualism*, 8(6), 776–780.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, 57(4), 421–457.

Studdert-Kennedy, M. (1998). The particulate origins of language generativity: from syllable to gesture. In Hurford, J., Studdert-Kennedy, M., & Knight, C., (eds.), *Approaches to the evolution of language*, pages 202–221. Cambridge University Press, Cambridge.

Svenonius, P. (2018). Learning rules versus learning items. *Linguistic Approaches to Bilingualism*, 8(6), 781–786.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, 57A(4), 745–765. https://doi.org/10.1080/02724980343000477

Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent-child interactions. *Cognition*, 148, 117–135. https://doi.org/10.1016/j.cognition.2015.11.002

Tyler, A. & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, 28(6), 649–667. https://doi.org/10.1016/0749-596X(89)90002-8

Visser, F. T. (1963). *An historical syntax of the English language*. Brill Archive.

Weir, M. W. (1964). Developmental changes in problem-solving strategies. *Psychological Review*, 71(6), 473. https://doi.org/10.1037/h0041785

White, L. & Genesee, F. (1996). How native is near-native? the issue of ultimate attainment in adult second language acquisition. *Second Language Research*, 12(3), 233–265. https://doi.org/10.1177/026765839601200301

Wittenberg, E. & Jackendoff, R. (2018). Formalist modeling and psychological reality. *Linguistic Approaches to Bilingualism*, 8(6), 787–791.

Xu, F. & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, 22(3), 531–556. https://doi.org/10.1017/S0305000900009946

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

Yang, C. (2013a). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16), 6324–6327. https://doi.org/10.1073/pnas.1216803110

Yang, C. (2013b). Who's afraid of George Kingsley Zipf? Or: Do children and chimps have language? *Significance*, 10(6), 29–34. https://doi.org/10.1111/j.1740-9713.2013.00708.x

Yang, C. (2015). Negative knowledge from positive evidence. *Language*, 91(4), 938–953. https://doi.org/10.1353/lan.2015.0054

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

Yang, C. (2017). Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*, 24(2), 100–125. https://doi.org/10.1080/10489223.2016.1274318

Yang, C. (2018). *A user's guide to the Tolerance Principle. Manuscript*. University of Pennsylvania (ling.auf.net/lingbuzz/004146).

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews*, 81(Part B), 103–119. https://doi.org/10.1016/j.neubiorev.2016.12.023

Yang, C. & Montrul, S. (2017). Learning datives: The tolerance principle in monolingual and bilingual acquisition. *Second Language Research*, 33(1), 119–144. https://doi.org/10.1177/0267658316673686

Yang, C. & Valian, V. (2018). Determiners and grammars. Submitted.

Yusa, N. (2018). Input effects on the development of I-language in L2 acquisition. *Linguistic Approaches to Bilingualism*, 8(6), 792–796.

Author's address

Charles Yang
Department of Linguistics and Computer Science
University of Pennsylvania
3401 Walnut Street 315C
Philadelphia, PA 19104, 215-898-7849

charles.yang@ling.upenn.edu