A User's Defense of the Tolerance Principle

Charles Yang
University of Pennsylvania

Feb 2023 (Final Version)

I was bemused by Prof. Dr. Hans-Olav Enger's recent article in this journal entitled "Type frequency is not the only factor that determines productivity, so the Tolerance Principle is not enough" (2022; henceforth E). Neither I, who is responsible for the said Principle, nor many others who have made use of it, ever held such opinions. From its conception some twenty years ago (Yang 2002b) to a comprehensive report in *The Price of Linguistic Productivity* (Yang 2016, henceforth POP) and its subsequent developments, the Tolerance Principle (TP) has always been an elucidation of the precise manner in which frequency interacts with other components of the grammar: it is never the only factor. It is even more surprising that the title does in fact reflect Enger's understanding of the matter: "according to the TP, minority rules in inflection should never become productive, at least not by a morphological path. This claim is not only strong, it is also flatly wrong" (E, p181). I wonder if we are talking about the same work.

High frequency as a prerequisite for productivity is an enduring insight that dates back to the structuralist tradition ("statistically predominant"; Nida 1949, p45) and also features prominently in generative grammar ("the ratio of possible to actually listed words"; Aronoff 1976, p36) and statistical accounts of language (e.g., the critical mass hypothesis; Rumelhart and McClelland 1986, Marchman and Bates 1994, Bybee 1995). However, simply equating high frequency with productivity immediately runs into problems. While statistical predominance does correctly establish the suffix *-ed* as the single productive process in the much-studied problem of English past tense, there are numerous cases where productivity can be found in non-dominant classes. For example, German noun pluralization uses several suffixes: *-s*, *-(e)n*, *-e*, *-er*, and the null *-ø*. Notably, the *-s* suffix covers the smallest number of nouns but is productively extended to novel words such as *iPhone-iPhones*. Indeed, multiple suffixes are productive as conditioned on gender and/or phonology of the noun and German-learning children overuse them, a true sign of productivity, throughout the early years of language acquisition (Elsen 2002, Kauschke et al. 2011). The German plural situation, then, can be accurately described as "local generalizations" according to Enger, who holds language as "a 'system' of low-level regularities, not all-encompassing rules" (E, p161). This is so because "speakers favour local solutions (small-scale generalizations)" (E, p184). But *favouring* is just a Panglossian restatement of the facts, not a theory: conflation of description with explanation runs throughout Enger's piece. In any case, we need to know why and when global rules are "favoured" in some cases (e.g., English past tense) while local ones are in others (e.g., German plurals).

It would have been a tremendous failure, as Enger charged, if the TP could not account for productive minority rules. In what follows, I will first describe the nature of the TP, which has been badly distorted by Enger. It will become immediately clear how minority rules can emerge as productive. I then discuss two cornerstone cases in Enger's discussion, English past tense and German plurals. Enger's misrepresentations further illustrate his failure to grasp the most elementary aspects of the TP as a formal model. Finally, I comment on other conceptual and empirical

questions Enger raises. The aim is to highlight the approach embodied by the TP: its mechanistic character enforces accountability, and in so doing provides a learning-theoretic account for patterns in language, perhaps at the expense of traditional approaches in linguistics.

⋆⋆

The TP is first and foremost a theory of learning. It specifies a precise threshold, as a proportion of items in the learner's experience, that a generalization can tolerate as exceptions: $\theta_N = N/\ln N$, where N is the cardinality of the item set.[1] Alternatively, we can think of the TP as a way to validate a function that forms a mapping between a domain and a co-domain. Critically, the TP holds that the rule or the mapping needn't be perfect: as long as it is good enough to cover a sufficiently large number of items, generalization is warranted. It is impossible for frequency to be the only factor: every TP-based frequency calculation pertains to a rule or function, which must be defined in terms of *something*.

In all domains of learning, we need to project a hypothesis from incomplete data. For example, if seven out of ten new species encountered on an island are herbivores, should you assume so for the other three as well as those yet-to-be-discovered (POP, p171)? The decision invariably hinges on the weight of supporting evidence: it is possible that a model like the TP is implicated in tasks beyond language acquisition. Even within the realm of language, the TP has been applied to phonology, morphology, and syntax: these domains have very different structural properties but the nature of learning is the same, i.e., generalization over items, possibly with exceptions. Enger's attempted dig at "the old generative axiom that grammar does not 'count' in the arithmetical sense" (E, p170) falls flat, for he fails to appreciate an elementary distinction in the study of computation: the representation of what is computed, and the algorithm by which the representation is manipulated. The same sorting algorithm can work for a deck of cards as well as a list of words. In the formal study of learning, all algorithms "count". The TP is an attempt to separate the formal aspect of rule learning (the algorithm that weighs evidence quantitatively) from the substance over which rules are formed (the representation over which evidence is accumulated), echoing a familiar distinction in the study of language universals (Chomsky 1965, Payne and Yang 2022).

When applied to language, the TP works over vocabulary items in the input over which potential rules are defined. Given a set of words, if there is a sufficiently dominant rule to account for them, then a global rule is learned: the Tolerable number of exceptions are lexically memorized, and nothing more needs to be done. If that turns out not to be case, the set is partitioned into subdivisions: in the case of language, along phonological, semantic, formal, and other linguistic or nonlinguistic dimensions (e.g., social hierarchy, which is necessary for honorific marking). The search for productivity proceeds recursively within. As a result, local productive rules may be discovered, with narrower and more restrictive conditions of application, and these may include rules defined over minority classes. If productivity cannot be identified even locally, then the learner resorts to lexical memorization and does not generalize beyond the input data: the target function may be partial, and that has to be discovered by the learner. This last case corresponds

[1] As suggested by a reviewer, it is worth emphasizing that the TP pertains strictly to types. The token frequency of a type in N plays no role, except in the trivial sense that a type with higher token frequency is more likely to be learned by children thereby more likely to participate in N for productivity calculation. The view contrasts with, for example, Baayen's productivity measures (2009) which do incorporate token frequencies. Those measures, however, are not capable of capturing structural factors in productivity ("local generalizations") as has been pointed out long ago (Van Marle 1992). Although not intended as accounts for language acquisition, they cannot capture the distributional patterns of child language either; see, e.g., Björnsdóttir (2021) for empirical tests.

to phenomena such as morphological gaps and related matters of ineffability, the topic of an entire chapter in POP (Chapter 5; see also Gorman and Yang 2019), as well as the absence of productivity in syntax, another chapter (Chapter 6). In other words, under the TP, the learner discovers global rules if possible, local rules if necessary, and no productive rules if all else fails.

The whole volume of POP is devoted to the dynamical process of rule formation. The issue of minority productivity is raised in the opening pages and a formal discussion is provided (POP, Section 3.5) before any of the empirical studies in the subsequent chapters. A schematic illustration of how local productivity emerges is reproduced in Figure 1, where outer circles represent less restrictive rules. When a rule defined over the entire set, e.g., the global rule R3 in Figure 1, fails to reach productivity, the learner can subdivide the words into subsets, marked by features (i.e., A, B), and search for (local) productivity continues with these subsets. For example, suppose there are 50 words: 20 follow R1, 20 follow R2, and 10 follow R3. Clearly no rule can be globally productive as all three have more exceptions than the tolerance threshold ($50/\ln 50 = 12$). However, as in Figure 1, the learner may discover that R1 is productive for words with features (+A) and R2 is productive for words with features (+B), each of which may have a tolerably low number of exceptions of their own. With 40 "exceptions" removed by these two local rules, R3 can be productive: the minority rule in the present example, as well as the *-s* suffix in the case of German plurals briefly summarized in the later part of this paper.

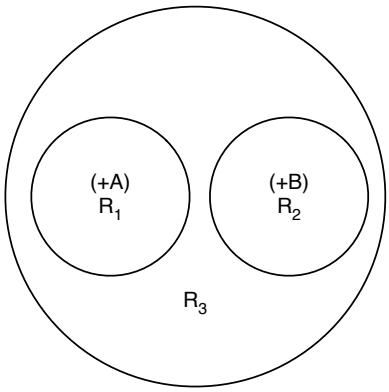

Figure 1: Recursive applications of the Tolerance Principle to detect structured rules (from POP, p74)

In POP, the divide-and-conquer search for productivity was carried out by hand over corpus data. Recent work has provided a full implementation (Belth et al. 2021). Subdivision focuses on one feature of the words at a time, drawing insight from the so-called one-dimensional-sort bias in learning and categorization (e.g., Medin et al. 1987). The Principle of Maximize Productivity (POP, p72) is invoked: if subdividing the words by feature A and by feature B both yields productive rules, the one with fewer exceptions is preferred. Since there is a finite number of features and a finite set of words, the search takes linear time. Furthermore, the Principle provides a way of picking out grammatically pertinent features on a language particular basis: they are those that lead to productive rules. Note that the Principle was proposed to help the learner make *local* decisions in the incremental process of learning: it is not, contrary to Enger's claim, a way to attain "the globally most efficient strategy" (E, p165). Again, the very point of the TP is that rule coverage only needs to be good enough.

I do not know how Enger could have missed these points: POP makes them abundantly clear in every case study. The English past tense, to which we return momentarily, is an outlier for

which a single global rule suffices. Yet the TP is "flatly wrong" (E, p181) because minority rules can be productive: Enger does not seem to understand the TP as an operational procedure at all. Opting to omit the mathematical details, Enger discusses the TP with an example: "a pattern of six members will not be productive, while a pattern of eight can" (E, p164). This misses the *proportional* nature of the TP: Enger identifies minority by absolute values. But the ratio of Tolerable exceptions ($1/\ln N$) is not a constant. Yes, 6 is smaller than 8 but the denominator also needs to be taken into account. A pattern that holds for 6 members will be productive if there are 10 members that are eligible for it ($10/\ln 10 = 4$), whereas a pattern of 8 will not be productive if the candidate set contains 14 items ($14/\ln 14 = 5$).

Before proceeding, let me stress that as a psychological theory of learning, the numerical values in the TP pertain to the learner's learning experience: in the case of language, the child's vocabulary. Children learn words incrementally and in fact quite slowly (Fenson et al. 1994, Bornstein et al. 2004). In some cases, adult-like productivity is established relatively late: the English past tense *-ed*, for example, is typically mastered by three years of age. This is because the irregular verbs in English are highly frequent and tend to be learned earlier: the *-ed* rule must gather enough regular verbs to overwhelm the irregulars. Only then do over-regularization errors emerge: prior to that point, children can only memorize verb-specific past tense forms and do not generalize beyond (Kuczaj 1977, Marcus et al. 1992, Yang 2002a, Tomasello 2003). Children who learn the rule earlier tend to be better/faster word learners, for whom the tolerance threshold is reached earlier (POP, Section 4.1.2). Moreover, productivity may fluctuate during the course of acquisition: the stochastic nature of child vocabulary acquisition may result in transient productivity of a rule that is unproductive in the adult language, only to correct itself after a larger and more representative vocabulary is acquired. The same holds for language change (POP: Section 5.2 and 5.3). The rise and fall of rule productivity in history are attributed to the changing vocabulary acquired by the language learners in the past. The cardinality values associated with a rule can change over time for both internal (e.g., independent phonological and semantic change, obscure words falling out of usage) and external (e.g., contact, borrowing, language policy) reasons, but the psychological calculus of learning for our ancestors is presumably the same as ours: Rule productivity changes as a consequence of changes in the input words, as we can see in the acquisition and change of English past tense.

⋆⋆⋆⋆⋆⋆⋆

According to the TP analysis (POP, Section 4.1), none of the irregular past tense forms in modern English is productive: every conceivable pattern has far too many exceptions. Even the most promising irregular class, the verbs that end in *ing* (/ɪŋ/), is splintered into several patterns that include vowel change to /æ/ (e.g., *sang*), vowel change to /ʌ/ (e.g., *swung*), idiosyncratic change (*brought*), and even the regular (e.g., *winged*): None emerges as the productive option under the TP. Only *-ed* rule is productive: the some 120 irregular verbs in routine circulation are clearly below the Tolerance threshold for the number of verbs that an English speaker knows. Consequently, the irregular patterns are not extended beyond those in the input while the regular pattern applies to both novel verbs (e.g., *googled*) and is often overused by children (e.g., *go-goed*, *hold-holded*) when their memorization of the irregular forms is not yet perfect.

The "dramatically different rates of overregularization and overirregularization" are not "debatable", contrary to Enger's assertion (E, p166). It is a long-standing observation in language acquisition, from Brown (1973) onward, that commission errors, which would include overirregularization

errors, are triflingly rare. Why not engage with the factual findings? Children over-regularize 8-10% of the time but over-irregularize only 0.2% of the time according to the most comprehensive study that evaluated 20,000 English past tense forms (Xu and Pinker 1995). A dense corpus study of a single child (Maslen et al. 2004) does not report a single over-irregularization error. Many conceivable analogical forms (e.g., *bite-bote*, *think-thunk*, *glow-glew*) are unattested, at least not in the some 3 million words of child English data in the public domain (MacWhinney 2000). The rates of overregularzation and overirregularization differ by two orders of magnitude: "dramatically different" by any standard. Such asymmetries hold in the cross-linguistic studies of child morphology extensively reviewed in POP (Chapter 3).

Puzzlingly, Enger endorses (E, p167) what Bybee and colleagues call process-oriented (more commonly known as source-oriented) and product-oriented forms (more commonly known as schemas). He does not seem to realize that the former is just another way of calling productive processes, where words meeting some structural description predictably follow some structural change (e.g., *-ed* applies to verbs), while the latter refers to unproductive ones, where words merely share some structural change by fiat (e.g., the very different-looking verbs that change rime to *ought* to form past tense). There is no problem in classifying rules as process- vs. product-oriented: Would the "dramatically different rates" between process- and product-oriented forms be more palatable? Alternative terminologies such as productive vs. unproductive (POP), regular vs. irregular (Xu and Pinker 1995), major vs. minor (Lightner 1968), phonological vs. morpholexical (Anderson 1974), etc. are all fine, as long as we understand them to be just that: Terminologies. In the words of a wise senior colleague, the real question is which rule wakes up in the morning and decides to be productive and which does not. The TP was developed in part to answer that question (Yang 2017).

Enger's discussion of English past tense conflates the synchronic and diachronic treatment of productivity. He points to the existence of doublets such as *dived-dove* and *sneaked-snuck* (Anderwald 2013, POP, Section 5.2) as a direct challenge to the conclusion that there is only one productive rule in English past tense. A form such as *dove* would be a problem if children spontaneously create them from *dive* as an instance of over-irregularization. But as reviewed above, they do not. Adults use *dove*: it is an irregular verb. Children will simply learn them as exceptions to the sole productive rule of *-ed*. The actual reason for the doublets is, of course, historical. As is well known, many currently irregular verbs are relics of once productive rules. Indeed, *snuck* may have been around in the English language all along albeit with low usage frequency but spread more broadly in certain varieties of American English (POP, p158f). A proper understanding of the past would involve an application of the TP to a reasonable approximation of the vocabulary set in the past.

Indeed, the diachronic application of the TP to "irregularized" verbs can be found in Ringe and Yang (2022). We focused on the history of the verbs with an /ɪ/ in the stem followed by a single velar consonant or homorganic velar cluster. Using the Penn-Helsinki Parsed Corpus of Early Modern English and the OED, which provided the first date of attestation, we applied the TP to the sets of such verbs whenever a new member became part of the English lexicon to calculate the productivity of rules at that particular time. We found that the ɪ→ʌ rule was productive well into Early Modern English, accounting for the pattern of *stick-stuck*, *string-strung*, *dig-dug*, etc. when these verbs were incorporated into English. The reason for this period of productivity is the fact that the past tense of *sing* and *ring* showed *sang* ∼ *sung* and *rang* ∼ *rung* variation (Taylor 1994): the two items *sung* and *rung* were able to swing the pendulum in the direction of productivity because the cardinality of this set of verbs was very small then (as is now). However, when *sung* and *rung* as past tense fell into disuse, the pendulum swung the other way and the rule became unproductive. Therefore, verbs such as *wing*, *pick*, and *rig*, which entered English in the 17th century and after, all took *-ed*

instead of becoming *wung*, *pug*, and *rug*. The verbs that actually follow the ɪ→ʌ rule — now an unproductive one — must be lexically memorized and thus prone to over-regularization by children (e.g., *digged*). Perhaps in the long run, some of the less frequent members will take on *-ed* instead. To these verbs we can add two more that became mainstream in the past twenty years or so: *bing* (Microsoft search engine) and *bling* (ostentatious display of jewelry or wealth). Both are uniformly inflected as regular, again affirming that none of the irregular processes is currently productive (which would have yielded *bang/bung* and *blang/blung*). Our analysis of change in productivity crucially relies on the assumption that the vocabulary of learners in the past can be reasonably approximated by historical corpora; see Kodner (2019, 2020) for extensive discussion and empirical validation.

The TP may also offer an account of genuine innovation as a result of rule misconvergence. I considered such a case in POP (Section 4.1.1). In the CHILDES database of child language (MacWhinney 2000), some children produced *brang* for the past tense of *bring* and *swang* for *swing*: interestingly, this is the only systematic pattern observed in the tiny number of over-irregularizations (Xu and Pinker 1995). These errors are predictable if the child has not mastered the productivity of the *-ed* rule but has acquired the verbs *bring-brought*, *sing-sang*, and *ring-rang*, all of which are very frequent and are likely learned quite early. At this transitory stage, the rule ɪ → æ / __ ŋ is productive as 2/3 is sufficient for generalization. Over time, and likely soon after, the rule will lose its productivity when children acquire a more complete vocabulary of English verbs. However, if they failed to acquire additional verbs in this class for whatever reason, the ɪ → æ / __ ŋ rule would productively spread: a recapitulation of history.

These few verbs, I hope, are sufficient to show the nature of productivity under the TP: it all comes down to the acquired vocabulary and the structural generalizations that can be established within. Since vocabulary acquisition is a stochastic process, productivity may rise and fall, including rules in the minority /ɪŋ/ class.

⋆

The study of German plurals gives a clear illustration for the recursive application of the TP. Recall that there are multiple suffixes with none anywhere near a global majority. The TP thus compels the learner to search for productive rules with subdivided sets of words. In POP (Section 4.4), I carried out, by hand, such an analysis for some 450 nouns that are representative of a young child's vocabulary. Not knowing German at all, my intention was to mimic the process of language acquisition: the grammar of German plurals must be discoverable by children. The analysis was successful: the plural suffixes are predictable, via multiple productive rules conditioned on phonology and/or gender, for over 80% of the nouns. A recent in-depth study of German noun plurals (Trommer 2021) comments on the POP analysis: "In fact, Yang (2016) has recently shown that *all* suffixation patterns of the German plural can be effectively learned in a principle-based computational model positing productive rules for patterns surpassing a general cognitive exceptionality threshold, basically extracting a subset of the generalizations assumed in this paper from corpus data" (p648, emphasis original). The implementation of the TP as a learning model (Belth et al. 2021) has automated the rule discovery process. The order in which the productive rules are identified from a small, child-appropriate, corpus closely mirrors that by German-learning children (Gawlitzek-Maiwald 1994, Elsen 2002, Kauschke et al. 2011). The induced rules also provide good coverage for nouns not used for learning (drawn from CELEX). Many generalizations previously noted in the German linguistics literature can be automatically identified, including

Wiese's "reduced final syllable" constraint (1999, p124) that words ending in a schwa followed by /l/, /r/, and /n/ add the null suffix -ø.

It would have been good for Enger to engage with this treatment of German plurals: it clearly shows that minority rule is not a problem. Instead, he takes issue with the status of noun gender in plural formation. In German, the plural form of a noun has been proposed as a way to determine the gender of the noun, an issue I was aware of and discussed (POP, p126-127). Enger challenges my analysis presumably because it used a list of nouns already annotated for gender: gender acquisition is thus a precondition for plural marking. He cites Müller's 2000 study ("the grammatical features gender and number are discovered simultaneously in language acquisition") to support the entangled nature of gender and plural. Unfortunately, this is a misreading of that work. Müller was concerned with the conceptual and grammatical expression of number (e.g., children's understanding of numerals, the misuse of a plural determiner on the singular), which is a prerequisite for the choice of plural suffixes but not the same thing (see Payne 2022 for how the child may determine which grammatical features require marking, a language specific matter). And the selection of Müller's paper is curious: her study is based on bilingual children acquiring French and German, an unnecessary complication, especially when there is a large monolingual acquisition literature available (see below).

The chicken-and-egg problem that Enger envisions may seem intractable to the linguist and their theories, but it won't concern the child equipped with a plausible psychological model of learning.

First, Enger overstates the mutual dependency between gender and plural. Gender is indeed used to predict the plural but only for a minority of nouns. In my analysis of 450 nouns, over 60% are fully accounted for phonology, in particular two rules: nouns ending in schwa take *-n* without exception, and nouns in Wiese's "reduced final syllable" class (see above) take -ø with a Tolerable number of exceptions. For the majority of nouns, then, gender is not needed at all for the plural. If only Enger bothered to check the final list of plural rules (POP, p133): only two out of six rules make reference to gender.

Second, noun gender can, and probably *is*, learned without making reference to the plural at all. In particular, the definite determiner in the nominative singular (*der*, *die*, *das*) provides unambiguous evidence, by far the most frequent and salient cue, and is the earliest determiner learned by children (Mills 1986). Even at 14 to 16 months of age – that is, at the very beginning of speech – German-learning infants already show knowledge of determiners as a formal category for grouping nouns (e.g., Höhle et al. 2004). As a result, it has been observed that "(N)oun gender in German is probably learnt via associating a particular lexical item with a gender marked determiner" (Szagun 2004, p27). It would be surprising if children were to opt for the plural as the cue for gender: statistically, plurals appear several times less frequently than singulars.

Finally, let us suppose, for the sake of argument, that the plural *can* help determine gender, perhaps at a later stage of language acquisition as I discussed in POP (p131, f22). Would that be a problem? Enger writes:

> A noun such as *Knauf* 'knob', with *-es* in the genitive singular and umlaut in the nominative plural, cannot be a feminine; more the 90% of such nouns are masculine. Of nouns inflecting like *Biene* 'bee', more than 90% will be feminine, according to Fedden/Corbett. Do we really want to exclude the possibility that children can observe this? (E, 169)

Of course we cannot, but the burden of proof is on Enger to demonstrate that children in fact do: his incredulousness is no proof. And if he succeeded in so doing, it would be a welcome result: the TP can surely pick out 90% rules as productive.

He continues:

> Fedden/Corbett also argue that, for the five largest inflection class in the German lexicon, the best prediction that can be made on the basis of gender accounts for less 30% of the data. The consequences for the TP should be clear. (E, 169-170)

Again, I have to question whether Enger read POP at all: phonology plays a bigger role in the prediction of the plural suffix than gender. Actually, Fedden/Corbett's number sounds about right. Phonology, as noted earlier, determines the plural suffix for over 60% of nouns, which leaves about 30% for gender. (Don't forget the exceptions which, according to the TP-based analysis, constitute about 12% of the data.) Even if gender had no predictive power on the plural (or any other inflection class), it would still not be a problem for the TP: A learner attempting to discover productive rules simply fails to succeed, and lexcialized memorization ensues. There are languages where gender assignment is idiosyncratic and must be learned by rote. Why should anyone be surprised by this possibility? Every word is an instance of Saussurean arbitrariness.

I do agree with Enger's statement: "Learners/speakers are generally opportunistic, clutching at every straw" (E, p168). Unfortunately, being "generally opportunistic" is not an explanation – no more than "favouring" (E, p184). How do learners/speakers know which straw to grasp at, or which one of the straws can prove productivity-saving? There is a book on that.

⋆⋆⋆⋆⋆⋆⋆⋆

Enger follows the discussion of English and German with a number of "counterexamples" that minority rules are productive and can spread to new members. I will not reiterate the logic for minority rules nor the procedure by which they can be identified. While I have no reason to doubt Enger's descriptions, they are of little use as they offer no workable basis for others to reanalyze these examples — unless you are intimately familiar with the Stavanger and North Gudbrandsdalen dialects of Norwegian, for instance. If Enger were serious about these cases as a challenge to the TP, he should have done the calculation to show that they are genuine counterexamples. This is the most disappointing aspect of his piece: a critique of an equation without calculating it a single time.

Enger's failure to provide data-driven analyses or the data for others to evaluate reflects a larger problem in descriptive and theoretical linguistics. While the issue of replicability has been a prominent theme in psychology and other behavioral and social sciences, it would be wise for linguistics to do some housekeeping of its own. Everyone can tell a good story about three verbs: How that story fares with other verbs should not be left for imagination. At the minimum, empirical claims ought to be accompanied with a dataset for verifiability. If computer scientists must make code accessible and biologists must submit DNA sequences and tissue samples, why shouldn't linguists make a list of words public? We would no longer just have to take each other's word for it.

Fortunately, the next generation of scholars is making linguistics a more accountable science. I will just mention a few recent studies: they all make use of the TP, which forces the researcher to be explicit about the data for otherwise calculation cannot be done. These include a study of gender assignment in Icelandic (Björnsdóttir 2021), noun diminutive suffixation in Dutch (van Tuijl and Coopmans 2021), argument structure mapping in English (Pearl and Sprouse 2021), verbal inflection variation and change in Frisian (Merkuur 2021), possessive suffix in Northern

East Cree (Henke 2022), the inflection of past participles in Latin (Kodner 2022), the changes in the English metrical stress system due to the influx of Latinate vocabulary (Dresher and Lahiri 2022), among others. These studies not only contain quantitative corpus results but also present converging evidence from experiments, naturalistic production, dialectal surveys, historical sources, etc. in support to their theoretical conclusions. Kodner, for instance, curated the largest dataset of Latin verbs to date: the poor coverage of previous theoretical accounts becomes blindingly obvious. Evidently we can't just take each other's word for it.

In my view, accountability is the most important feature of the TP. It makes a crisp prediction about productivity whereas other accounts make gestures toward an essentially infinite list of factors without articulating how they interact; see Yang (2015) on one such proposal that is based on, yes, frequencies. Being a parameter-free model, the TP enables researchers to make unambiguous predictions including targeted manipulations in artificial language experiments (Schuler 2017). For instance, Emond and Shi (2021) 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 permutation; in the second, 10 out of the 16 items followed a word order permutation. The design reflects the critical prediction of the TP. For $N = 16$ items, the critical threshold is $16/\ln 16 = 5.77$: 10 is insufficient for generalization despite being the majority but 11 is. In fact, 14-month-old infants generalized the pattern from the 11/16 set, but not the 10/16 set.

Enger regurgitates "conceptual" critiques of the TP (E, p170-171) by several commentators including Wittenberg and Jackendoff, Kapatsinski, and Goldberg; see Yang 2018 for response. They found it difficult to reconcile the assumption of serial processing that leads to the derivation of the TP with the "fact" that the brain is "parallel". But they make no mention of the vast behavioral evidence for serial effects in many domains including numerosity, memory, and indeed language processing. A more responsible take would be to provide a theory of how a parallel brain can yield serial behavior – reciting the digits of π, for example – instead of refusing to engage with the empirical results. Conceptual arguments are cheap especially when grounded in ignorance: Do these critics have any idea how a supposedly parallel brain produces *parallel* effects?

It is fair to ask deeper questions about the psychological and neurological underpinnings of the TP, as one would for any theory of learning. For example, how do infants process quantities like 10, 11, and 16, surely unconsciously, such that the very small difference of 1 results in qualitatively different results? At the same time, we shouldn't be too surprised that they could: even ants have a discrete counter (Wittlinger et al. 2006). Nevertheless, the cognitive capacity for tracking proportional statistics has been amply demonstrated by the robust finding of transitional probability learning in language and other domains (Saffran et al. 1996). In fact, the TP is computationally simpler: it only needs to track a proportion of types, rather than a proportion of toke frequencies (of types) as in transitional probability learning. Finally, I have discussed the unreasonable effectiveness of the TP (POP, p76f.) and conclude the volume with the confession "... the most important conclusion from the present study is not whether the Tolerance Principle is ultimately correct. It is much more important that something like the Tolerance Principle can be established in the first place; by working out the axioms of language and cognition to their deductive ends — which is why it had to be in the form of an equation. This still strikes me as the most exciting aspect of generative grammar, even if the solution on offer turned out to be a lucky guess" (p227). The TP should be challenged and better understood but the assessment should be fair, accurate, and ultimately empirical.

⋆⋆

Enger's misunderstanding of the TP partly stems from his lack of familiarity with the mathematical and developmental study of language. But I suspect that there is a more fundamental disconnect.

As a learning model, the TP is a *discovery procedure* (Chomsky 1957): given a corpus as input, it produces a grammar as output, as in the implementation of Belth et al. (2021) and subsequent work. The learning-theoretic nature of the TP, and its likely domain generality, in effect pit the child against the linguist; or more bluntly, the TP against other theoretical devices. The task is all the same: to identify significant generalizations that reside in a finite sample of data. The linguist should discover exactly the same generalizations as the child: no fewer but no more either, if the goal is to understand our biological capacity for language.

Suppose the linguist's favorite toolkit – a morphological principle such as the No Blur Principle (Carstairs-McCarthy 1994) that Enger appeals to (E, p177), some typological generalization, or just a hunch – led to some discovery. If that discovery can be replicated by a TP-like procedure, then the linguist's toolkit becomes redundant and should be dispensed with: What can be learned needn't be built in. It is probably worth mentioning that I started the work on the TP some twenty years ago when I found the locality constraints in Distributed Morphology designed to capture productivity (e.g., Marantz 2001) to be inadequate. The TP shows that the solution for productivity does not lie in the architecture of the grammar but in the mechanism of learning. Similarly, while I assumed the innateness of many syntactic parameters in my earlier work (Yang 2002a), I no longer believe that is necessary: the word order variation across languages can be learned by the TP over lexical items (POP, Yang et al. 2017). More positively, a discovery procedure can be viewed as a useful baseline: theoretical assumptions are strengthened if they capture generalizations that elude a discovery procedure.

It is useful to stress that results discovered from corpora, even when strictly speaking true, may not be explanatorily revealing. For example, one could easily put frequency, phonology, and gender of German nouns into a multivariate regression model to predict the plural suffix. All factors may come out statistically significant but the precise manner in which they interact would still remain obscure – except maybe the test finds that certain factors "interact" in a statistical sense. We would only know what matters but not how. Here the linguist still has the advantage. While there is no guarantee that every generalization we propose is correct, at least we can make our theories mechanistically interpretable, which is no longer a given in the age of big data and bigger machines. We need independent and converging evidence from multiple sources to hold descriptive statements accountable, whether they are generated by linguists or statistical packages. Coverage of data, especially data beyond those used to form generalizations, is of paramount importance. After all, children generalize beyond the input when they learn a language.

It would be good for Enger, a specialist on gender and inflection, to engage with the TP in a productive fashion. As a formal category, gender may be distributionally dependent on semantic, phonological, morphological, and other attributes of the noun. And the direction of dependency needn't be uniform even within a single language (Enger 2004). One indeed needs to be opportunistic, as with the German plurals: some are predictable from phonology, some from gender, some from a combination of both, and yet others must be memorized by rote. All aspects of language work like this: a cocktail of systematicity and arbitrariness. We need explicit models to tease them apart.

I encourage Enger to take a corpus and give the TP a try. If a procedure can be implemented on a computer — the code for Belth et al. (2021) is on GitHub — it should be sufficiently mechanical for anyone to follow. Doing so, however, does require a careful reading of the manual.

References

Anderson, Stephen R. 1974. *The organization of phonology*. New York: Academic Press.

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

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Baayen, R Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In *Corpus linguistics. an international handbook*, ed. Anke Ludeling and Merja Kyto, 900–919. Mouton de Gruyter.

Belth, Caleb, Sarah Payne, Deniz Beser, Jordan Kodner, and Charles Yang. 2021. The greedy and recursive search for morphological productivity. In *Proceedings of CogSci 2021*.

Björnsdóttir, Sigríur Mjöll. 2021. Productivity and the acquisition of gender. *Journal of Child Language* 48:1209–1234.

Bornstein, Marc H, Linda R Cote, Sharone Maital, Kathleen Painter, Sung-Yun Park, Liliana Pascual, Marie-Germaine Pêcheux, Josette Ruel, Paola Venuti, and Andre Vyt. 2004. Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development* 75:1115–1139.

Brown, Roger. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Bybee, Joan L. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10:425–455.

Carstairs-McCarthy, Andrew. 1994. Inflection classes, gender, and the principle of contrast. *Language* 70:737–788.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Dresher, B. Elan, and Aditi Lahiri. 2022. The foot in the history of English: Challenges to metrical coherence. In *English historical linguistics: Change in structure and meaning*, ed. Bettelou Los, Claire Cowie, Patrick Honeybone, and Graeme Trousdale, 42–59. John Benjamins.

Elsen, Hilke. 2002. The acquisition of German plurals. In *Morphology 2000: Selected Papers from the 9th Morphology Meeting, Vienna, 25-27 February 2000*, volume 218, 117. John Benjamins Publishing.

Emond, Emeryse, and Rushen Shi. 2021. Infants' rule generalization is governed by the Tolerance Principle. In *Proceedings of the 45nd annual Boston University Conference on Language Development*, ed. Danielle Dionne and Lee-Ann Vidal Covas, 191–204.

Enger, Hans-Olav. 2004. On the relation between gender and declension: A diachronic perspective from norwegian. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 28:51–82.

Enger, Hans-Olav. 2022. Type frequency is not the only factor that determines productivity, so the Tolerance Principle is not enough. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 144:161–187.

Fenson, Larry, Philip S Dale, J Steven Reznick, Elizabeth Bates, Donna J Thal, Stephen J Pethick, Michael Tomasello, Carolyn B Mervis, and Joan Stiles. 1994. Variability in early communicative development. *Monographs of the Society for Research in Child Development* i–185.

Gawlitzek-Maiwald, Ira. 1994. How do children cope with variation in the input? The case of German plurals and compounding. In *How tolerant is Universal Grammar? essays on language learnability and language variation*, ed. Rosemarie Tracy and Elsa Lattey, 225–266. Tübingen: Niemeyer.

Gorman, Kyle, and Charles Yang. 2019. When nobody wins. In *Competition in inflection and word formation*, ed. Franz Rainer, Francesco Gardani, Hans C. Luschützky, and Wolfgang U. Dressler, 169–193. Berlin: Springer.

Henke, Ryan E. 2022. Rules and exceptions: A Tolerance Principle account of the possessive suffix in Northern East Cree. *Journal of Child Language* 1–36.

Höhle, Barbara, Jürgen Weissenborn, Dorothea Kiefer, Antje Schulz, and Michaela Schmitz. 2004. Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy* 5:341–353.

Kauschke, Christina, Anna Kurth, and Ulrike Domahs. 2011. Acquisition of German noun plurals in typically developing children and children with specific language impairment. *Child Development Research* 2011.

Kodner, Jordan. 2019. Estimating child linguistic experience from historical corpora. *Glossa* 4:122.

Kodner, Jordan. 2020. Language acquisition in the past. Doctoral Dissertation, University of Pennsylvania, Philadelphia.

Kodner, Jordan. 2022. Language acquisition guiding theory and diachrony: A case study from Latin morphology. *Natural Language and Linguistic Theory* 1–60.

Kuczaj, Stan A. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior* 16:589–600.

Lightner, Theodore M. 1968. On the use of minor rules in Russian phonology. *Journal of Linguistics* 4:69–72.

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

Marantz, Alec. 2001. Words. In *20th West Coast Conference on Formal Linguistics*. University of Southern California.

Marchman, Virginia A., and Elizabeth Bates. 1994. Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language* 21:339–366.

Marcus, Gary, Steven Pinker, Michael T. Ullman, Michelle Hollander, John Rosen, and Fei Xu. 1992. *Overregularization in language acquisition*. Monographs of the Society for Research in Child Development. Chicago: University of Chicago Press.

Maslen, Robert, Anna L. Theakston, Elena V.M. Lieven, and Michael Tomasello. 2004. A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language and Hearing Research* 47:1319–1333.

Medin, Douglas L, William D Wattenmaker, and Sarah E Hampson. 1987. Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology* 19:242–279.

Merkuur, Anne. 2021. Changes in modern Frisian verbal inflection. Doctoral Dissertation, University of Amsterdam.

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

Müller, Natascha. 2000. Gender and number in acquisition. In *Gender in grammar and cognition*, ed. Barbara Unterbeck, 351–400. Mouton de Gruyter.

Nida, Eugene A. 1949. *Morphology: The descriptive analysis of words*. Ann Arbor: University of Michigan Press, 2nd edition.

Payne, Sarah. 2022. When collisions are a good thing: the acquisition of morphological marking. Bachelor's thesis, University of Pennsylvania.

Payne, Sarah, and Charles Yang. 2022. Making good on BADS: Commentary on Dressler et al. 2020. *Italian Journal of Linguistics* In press.

Pearl, Lisa, and Jon Sprouse. 2021. The acquisition of linking theories: A tolerance and sufficiency principle approach to deriving utah and rutah. *Language Acquisition* 28:294–325.

Ringe, Don, and Charles Yang. 2022. The threshold of productivity and the irregularization of verbs in Early Modern English. In *English historical linguistics: Change in structure and meaning*, ed. Bettelou Los, Claire Cowie, Patrick Honeybone, and Graeme Trousdale, 91–111. Amsterdam: John Benjamins.

Rumelhart, David E., and James L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel distributed processing: Explorations into the microstructure of cognition. volume 2: Psychological and biological models*, ed. James L. McClelland, David E. Rumelhart, and the PDP Research Group, 216–271. Cambridge, MA: MIT Press.

Saffran, Jenny R., Richard N. Aslin, and Elissa Newport. 1996. Statistical learning by by 8-month-old infants. *Science* 274:1926–1928.

Schuler, Kathryn. 2017. The acquisition of productive rules in child and adult language learners. Doctoral Dissertation, Georgetown University, Washington, D.C.

Szagun, Gisela. 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* 31:1–30.

Taylor, Ann. 1994. Variation in past tense formation in the history of English. In *Penn working papers in linguistics 1*, ed. Roumyana Izvorski, Miriam Meyerhoff, Bill Reynolds, and Victoria Tredinnick, 143–158. Philadelphia: Penn Linguistics Club.

Tomasello, Michael. 2003. *Constructing a language*. Cambridge, MA: Harvard University Press.

Trommer, Jochen. 2021. The subsegmental structure of German plural allomorphy. *Natural Language and Linguistic Theory* 39:601–656.

van Tuijl, Rosita, and Peter Coopmans. 2021. The productivity of Dutch diminutives. *Linguistics in the Netherlands* 38:128–143.

Van Marle, Jaap. 1992. The relationship between morphological productivity and frequency: A comment on Baayen's performance-oriented conception of morphological productivity. In *Yearbook of morphology 1991*, ed. Geert Booij and Jaap van Marle, 151–163. Amsterdam: Springer Netherlands. URL `http://dx.doi.org/10.1007/978-94-011-2516-1`$_9$.

Wiese, Richard. 1999. On default rules and other rules. *Behavioral and Brain Sciences* 22:1043–1044.

Wittlinger, Matthias, Rüdiger Wehner, and Harald Wolf. 2006. The ant odometer: Stepping on stilts and stumps. *Science* 312:1965–1967.

Xu, Fei, and Steven Pinker. 1995. Weird past tense forms. *Journal of Child Language* 22:531–556.

Yang, Charles. 2002a. *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, Charles. 2002b. A principle of word storage. https://bit.ly/3OrnAT5https://bit.ly/3OrnAT5. Yale University Manuscript.

Yang, Charles. 2015. For and against frequencies. *Journal of Child Language* 42:287–293.

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

Yang, Charles. 2017. How to wake irregular (and speechless). In *On looking into words (and beyond): Structures, relations, analyses*, ed. Claire Bowern, Laurence Horn, and Raffaela Zanuttini, 211-232. Language Science Press.

Yang, Charles. 2018. A formalist perspective on language acquisition (Target article with peer commentary and author's reply). *Linguistic Approaches to Bilingualism* 8:665–809.

Yang, Charles, Stephen Crain, Robert C. Berwick, Noam Chomsky, and Johan J. Bolhuis. 2017. The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews* 81:103 – 119.