A User's Guide to the Tolerance Principle

Charles Yang
charles.yang@ling.upenn.edu

July 2018

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, is

$$\sum_{r=1}^{N} r \frac{1}{rH_N} = \frac{N}{H_N}$$

Similar calculation can be made for a productive rule with e exceptions that must be listed before the application of the rule. The resulting complexity, which contains the terms N and e, is compared against N/H_N above. This yields the solution $\theta_N = N/\ln N$: if e is no greater than θ_N, then it is more efficient to have a productive rule. The solution makes use of the approximation $H_N \approx \ln N$, the focal point of Section 4 below.

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.

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

N	θ_N	%
10	4	40.0
20	7	35.0
50	13	26.0
100	22	22.0
200	38	19.0
500	80	16.0
1,000	145	14.5
5,000	587	11.7

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.

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 1 provides the vocabulary size estimates of American English-learning children from a large scale study (Hart and Risley 1995).

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-

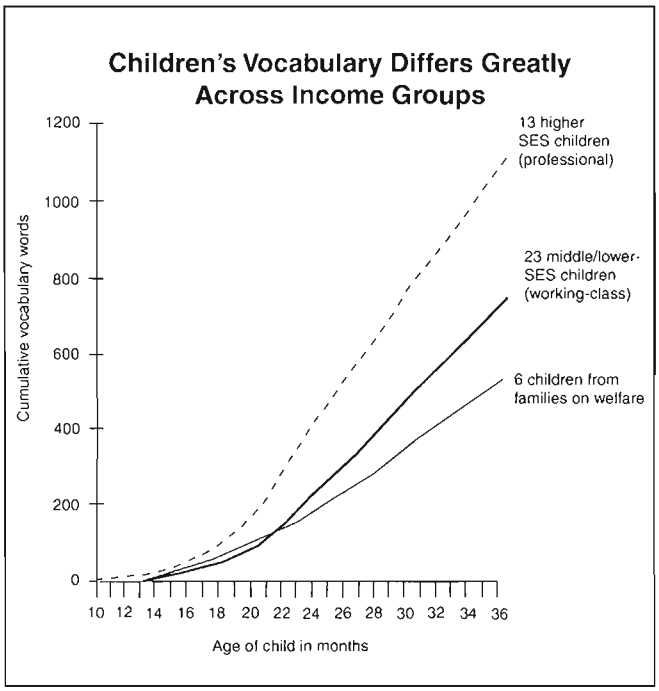

Figure 1: Vocabulary size estimates of American children.

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 1. 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 *-ø*). 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.

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%

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 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, as can be been in Table 4. 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. Finally, Rushen Shi (personal communication) informed me of a similar artificial language experiment (Koulaguina and Shi 2017). In that artificial language, words are introduced to learners 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 that children just keep track of two quantities. It is virtually certain that these quantities cannot be

Table 4: The Harmonic Number and the tolerance threshold.

N	H_N	$\ln N$	θ_N with H_N	θ_N with $\ln N$
2	1.500000	0.693147	1	2
3	1.833333	1.098612	1	2
4	2.083333	1.386294	1	2
5	2.283334	1.609438	2	3
6	2.450000	1.791759	2	3
7	2.592857	1.945910	2	3
8	2.717857	2.079442	2	3
9	2.828969	2.197225	3	4
10	2.928968	2.302585	3	4
15	3.318229	2.708050	4	5
20	3.597740	2.995732	5	6
50	4.499206	3.912023	12	13
100	5.187378	4.605170	20	22
200	5.878032	5.298317	35	38
500	6.792824	6.214608	74	81
1000	7.485478	6.907755	134	145

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.

Hence my final suggestion for the Tolerance Principle: Use it until it breaks.

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.

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*, page In press. Springer.

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. (2017). Rule generalization from inconsistent input in early infancy. Unpublished manuscript, UQAM.

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 Lingustics 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 *Festschrift for XXX*. Forthcoming.

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.