

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Intolerant Data: Testing The Tolerance Principle

Permalink

<https://escholarship.org/uc/item/9f75m0p4>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors

Ma, Xiaomeng

Chodorow, Martin

Valian, Virginia

Publication Date

2022

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Intolerant Data: Testing The Tolerance Principle

Xiaomeng Ma (xma3@gradcenter.cuny.edu)

Department of Linguistics, 365 5th Ave
New York, NY 10016 USA

Martin Chodorow (martin.chodorow@hunter.cuny.edu)

Department of Psychology, 695 Park Ave
New York, NY 10065 USA

Virginia Valian (virginia.valian@hunter.cuny.edu)

Department of Psychology, 695 Park Ave
New York, NY 10065

Abstract

Rule-based learning is an important aspect of language acquisition. Yang (2005, 2016) proposed the Tolerance Principle (TP) to predict when a rule will be formed by the language learner. We present the derivation of the TP as originally proposed and test it on both hypothetical data and corpus data from 8 children. Results for the hypothetical data contradict the TP's predictions, as do the data from 7 of the 8 children. We conclude that the original form of the TP does not explain rule-learning.

Keywords: Tolerance Principle; Corpus Analysis; Past Tense Overregularization

Introduction

Rule-based learning, such as past-tense acquisition, is an important feature of language acquisition. For example, children form the regular past tense for a novel verb in an experimental setting, e.g. *wug* - *wugged* (Berko, 1958) and make overregularization errors on irregular verbs, e.g. *hold* - **holded* (Marcus et al., 1992; Pinker & Ullman, 2002). Such evidence indicates that the rule is productive since the rule produces word forms that children have not previously encountered in their input. But what leads to development of rules in the first place?

Yang (2005, 2016) has proposed the Tolerance Principle (TP) to predict when a productive rule will be deployed by the language learner. The cognitive motivation for rule-generation is a reduction in lexical access time. According to Yang's derivation, lexical access time depends on the number of exceptions (e) in the data. When the number of exceptions (e) is smaller than the total number of items (N) divided by the natural log of N , $e \leq \frac{N}{\ln N}$, the rule will be deployed. Experimental studies of rule-learning with both children and adults have supported the predictions of the TP (Schuler, Yang, & Newport, 2021; Emond & Shi, 2021).

Our studies do not address the cognitive motivation of the TP, but test the predictions of the original formulation. We first present the derivation of the TP. We then use the formula to test hypothetical data and children's corpus data. We conclude that the original formulation yields unexpected inconsistencies and does not fully account for children's first (detected) over-regularization errors.

Deriving the Tolerance Principle

First Steps

A productive rule will be deployed when it delivers a more efficient result than not using a rule. Yang (2016) used lexical access time to measure efficiency and hypothesized that a productive rule will reduce the average time required to retrieve the target form, e.g., the past tense form. He proposed two different models, one for rule-based processing and one for no-rule processing. A rule will be used when the time complexity of rule-based processing is smaller than that of no-rule processing.

In the no-rule model, all the lexical forms (e.g., the past tense forms) are retrieved from memory using a serial search process (Forster, 1976, 1992). The lexical items are stored in a ranked list based on their frequency, with the most frequent items at the top. In order to retrieve an item at position i , the model sequentially searches all the $i-1$ items ranked higher than i until it reaches position i . The average time complexity (T) for the whole list is calculated as the sum of the each word's lexical access time (t_i) multiplied by its probability (p_i), $T = \sum_{i=1}^N (p_i \cdot t_i)$.

The t_i is approximated as its rank, $t_i = r_i$ (Murray & Forster, 2004).¹ The p_i is approximated based on the assumption that the frequencies of the items in the ranked list follow a Zipfian distribution (Zipf, 1949). In a Zipfian distribution, the product of a word's frequency (f_i) and its rank (r_i) is a constant C , i.e. $f_i \cdot r_i = C$. By replacing f_i with $\frac{C}{r_i}$, p_i can be expressed

$$\text{as: } p_i = \frac{f_i}{\sum_{k=1}^N f_k} = \frac{\frac{C}{r_i}}{\sum_{k=1}^N \frac{C}{r_k}} = \frac{\frac{1}{r_i}}{\sum_{k=1}^N \frac{1}{r_k}}.$$

Therefore, by substituting for p_i and t_i , the average time complexity for the no-rule model (T_{NoRule}) for the list can be expressed as: $T_{NoRule} = \sum_{i=1}^N \left(\frac{1}{\sum_{k=1}^N \frac{1}{r_k}} \cdot \frac{1}{r_i} \right)$. This value, $\sum_{k=1}^N \frac{1}{r_k}$, is the Harmonic number H_N , so $T_{NoRule} = \frac{N}{H_N}$.

¹Yang (2018) simplified Murray and Forster (2004)'s rank hypothesis as 'the i -th ranked item takes i units of time to be retrieved'.

The rule-based model adds an implementation of the Elsewhere Condition (Anderson, 1969; Halle & Marantz, 1993) to the no-rule model. When a productive rule is deployed, all the items are divided into rule-based items and exceptions. The exceptions are retrieved from memory and stored in a frequency-ranked list, as in the no-rule model.

The rule-based items are generated by a rule-applying process and concatenated into an unordered set at the end of the exceptions list. To access the target form of an item w_i , the model first sequentially searches the list of exceptions for a match for w_i . If w_i is found, meaning that w_i is an exception (e.g. *go*), the target form (*went*) is retrieved. If, however, w_i is not found through an exhaustive search of the exceptions list, because w_i is a rule-based item (e.g. *want*), the rule is applied to w_i (*wanted*). In sum, the rule is applied after an exhaustive search of the exceptions. As the number of exceptions increases, the time to apply the rule will increase.

Calculation of the average time complexity of the rule-based model consists of two parts, the T_E for the exceptions list, which is the same as the no-rule model, and T_R for the rule-based set. Yang assumed that the exceptions also follow a Zipfian distribution, like all the items in a list. Therefore, T_E is calculated by substituting the number of exceptions e for N , $T_E = \frac{e}{H_e}$. Since the rule-based items are reached after a complete search of the e exceptions, the time complexity for all the rule-based items is $T_R = e$. The overall time complexity of the rule-based model is the weighted sum of T_E and T_R , $T_{Rule} = \frac{e}{N} \cdot \frac{e}{H_e} + (1 - \frac{e}{N}) \cdot e$.

A productive rule will be deployed when $T_{Rule} \leq T_{NoRule}$. The maximum value of e is derived by solving this inequation:

$$\frac{e}{N} \cdot \frac{e}{H_e} + (1 - \frac{e}{N}) \cdot e \leq \frac{N}{H_N} \quad (1)$$

Approximation and Simplification

To solve the inequation above, Yang first approximated the Harmonic number with the natural log $\ln$, i.e. $H_N \approx \ln N$, $H_e \approx \ln e$. Inequation (1) thus becomes (2):

$$\frac{e}{N} \cdot \frac{e}{\ln e} + (1 - \frac{e}{N}) \cdot e \leq \frac{N}{\ln N} \quad (2)$$

Further, Yang solved inequation (2) by substituting $\frac{e}{N}$ with x and treating inequation (2) as a function of x :

$$f(x) = x \cdot \frac{e}{\ln e} + (1 - x) \cdot e - \frac{N}{\ln N} \quad (3)$$

$$= x^2 \left(\frac{1}{\ln N + \ln x} \right) + (1 - x) \cdot x - \frac{1}{\ln N} \quad (4)$$

It is observed that when $x = \frac{1}{\ln N}$, $f(x) \approx 0$ for large N . Therefore, by substituting $x = \frac{1}{\ln N} = \frac{e}{N}$, the results of the inequation (2) is simplified as $e \leq \frac{N}{\ln N}$.

After approximation and simplification, the Tolerance Principle is stated as:

Let R be a rule application to N items, of which e are exceptions. R is productive if and only if $e \leq \theta_N$, where $\theta = \frac{N}{\ln N}$ (Yang, 2016).

We examine the effects of the approximation and simplification of the inequation on the model output. As noted by Yang (2016), the approximation only works for large N . For small N s, the differences might be substantial. For example, when $N = 20$, $e \leq 20/\ln 20$ (6.8), $e \leq 20/H_{20}$ (5.6), and $x = 0.43$ for $f(x)$ in equation (3), which yields $e \leq 8.6$. In addition, the approximated results $e \leq N/\ln N$ actually simplified the inequation (1) from a quadratic function to a linear function², which might be inappropriate in numerical calculations.

Our first experiment tests the effects of the approximation and simplification across three sample sizes: $N = 10, 100$ and $1,000$. We solve inequation (1) to calculate the numerical θ and we compare its predictions with the approximated θ using $N/\ln N$. As summarized in Table 1, as N increases, the numerical difference between the approximated $\theta = N/\ln(N)$ and the numerical θ remains. We thus question whether $N/\ln N$ is a proper approximation of the maximum number of exceptions, since even a large sample shows a difference in the number of allowed exceptions.

Figures 1a - 1c plot the T_{Rule} and T_{NoRule} for different numbers of exceptions. As Fig. 1 shows, the function of T_{Rule} is quadratic with two solutions, $e \leq \theta$ or $e \geq N$. When $e = N$, Rule and NoRule are identical processes, and $e > N$ is not possible.

Table 1: Approximated θ and Numerical θ for different N s

N	Approximated θ $\theta = \frac{N}{\ln N}$	Numerical θ by solving $\frac{\theta}{N} \cdot \frac{\theta}{H_\theta} + (1 - \frac{\theta}{N}) \cdot \theta = \frac{N}{H_N}$
10	4.34	4.53
100	21.71	23.24
1,000	144.76	152.77

The Effects of Rank Permutation

In the derivation of the TP, time complexity was first calculated based on each item's probability and rank. The probability itself is approximated with frequency and rank ($f_i \cdot r_i = C$), on the assumption that the items (including exceptions) follow a Zipfian distribution. This approximation eliminates rank as a variable in the formula. Although all the items in a no-rule list would be expected to follow a Zipfian distribution, the same may not hold for the exceptions, especially for a data set with a small N . However, the time complexity of the exceptions $T_E \approx \frac{e}{H_e}$ in the model is derived under the assumption that the exceptions also follow a Zipfian distribution.

In order to test if rank affects the calculation of time complexity, we calculated T_{Rule} and T_{NoRule} using the probability

²In Yang(2016) page 63, he noted and plotted the quadratic function for inequation (1).

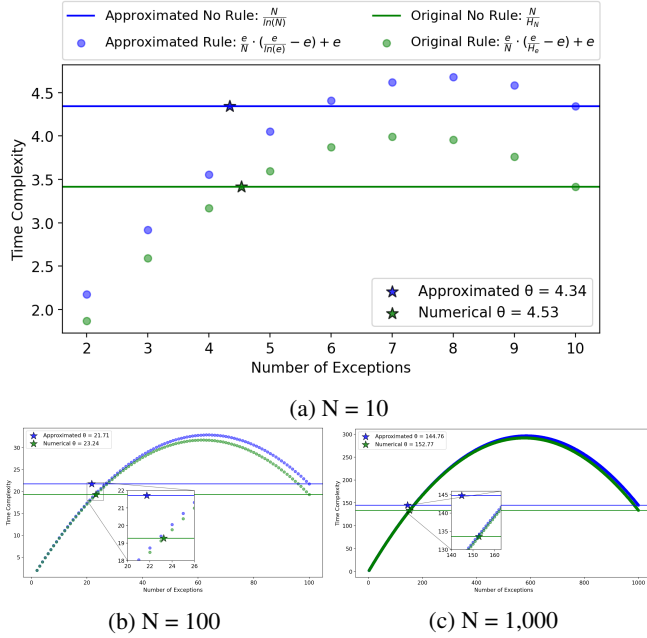


Figure 1: The plot of the time complexity for different values of N and e with $\ln N$ and H_N

and rank to determine when $T_{Rule} < T_{NoRule}$. We created a hypothetical list of 10 items whose frequencies follow a Zipfian distribution, with the highest frequency as 100 and the lowest as 10. In the baseline scenario all the items are exceptions; thus, no rule can be derived. The total average time complexity is $T_{NoRule} = 3.42$, as calculated in Table 2.

Table 2: Baseline Scenario: 10 exceptions, 0 regulars

Item	Frequency	rank	Time Complexity $T = \sum_{i=1}^N (p_i \cdot t_i)$
Excep.	100	1	$0.34 = 100/293 \times 1$
Excep.	50	2	$0.34 = 50/293 \times 2$
Excep.	33	3	$0.34 = 33/293 \times 3$
Excep.	25	4	$0.34 = 25/293 \times 4$
Excep.	20	5	$0.34 = 20/293 \times 5$
Excep.	17	6	$0.35 = 17/293 \times 6$
Excep.	14	7	$0.33 = 14/293 \times 7$
Excep.	13	8	$0.35 = 13/293 \times 8$
Excep.	11	9	$0.34 = 11/293 \times 9$
Excep.	10	10	$0.34 = 10/293 \times 10$
Total	293		3.42

Based on the TP's prediction, the approximated $\theta = 4.34$, and the numerical $\theta = 4.53$, meaning that the rule should be derived only if there are 4 or fewer exceptions. In the first experimental scenario, we randomly placed 7 exceptions and 3 regulars (ranked 2nd, 5th and 6th) in the list. Since $7 > 4$, no rule should be derived: $T_{Rule} > T_{NoRule}$. The $T_{NoRule} = 3.42$ is the same as the baseline scenario. However, $T_{Rule} = 3.29 < 3.42$, as calculated in Table 3, which is inconsistent with TP's prediction, whether we use approximated θ or numerical θ .

The frequency rank of the regulars substantially affects the value of T_{Rule} . In the second experimental scenario, if the

Table 3: 7 exceptions, 3 regulars ranked 2nd, 5th and 6th

No Rule			
Item	Frequency	rank	Time Complexity
Excep.	100	1	0.34
Regular	50	2	0.34
Excep.	33	3	0.34
Excep.	25	4	0.34
Regular	20	5	0.34
Regular	17	6	0.35
Excep.	14	7	0.33
Excep.	13	8	0.35
Excep.	11	9	0.34
Excep.	10	10	0.34
Total	293		T_{NoRule} 3.42
With a Rule			
Excep.	Freq.	rank	Time Complexity $T_E = \sum_{i=1}^N (p_i \cdot t_i) \cdot \frac{e}{N}$
Excep.	100	1	$0.24 = 100/293 \times 1 \times 0.7$
Excep.	33	2	$0.16 = 33/293 \times 2 \times 0.7$
Excep.	25	3	$0.18 = 25/293 \times 3 \times 0.7$
Excep.	14	4	$0.13 = 14/293 \times 4 \times 0.7$
Excep.	13	5	$0.16 = 13/293 \times 5 \times 0.7$
Excep.	11	6	$0.16 = 11/293 \times 6 \times 0.7$
Excep.	10	7	$0.17 = 10/293 \times 7 \times 0.7$
Total			1.19
Regular			Time Complexity
Regular	50		$T_R = e \cdot (1 - \frac{e}{N})$
Regular	20		
Regular	17		
Total			2.1 = 7×0.3
T_{Rule}			3.29 = $1.19 + 2.1 < 3.42$

three regulars are ranked 8th, 9th and 10th and $T_{Rule} = 3.77 > 3.42$, as calculated in Table 4, the results confirm the TP's prediction. Depending on the frequency rank of the regulars, the TP's prediction will be confirmed or disconfirmed in a dataset with the same number of exceptions.

Table 4: 7 exceptions 3 regulars ranked 8nd, 9th and 10th

No Rule			
Verb	Frequency	rank	Time Complexity
Excep.	100	1	0.34
Excep.	50	2	0.34
Excep.	33	3	0.34
Excep.	25	4	0.34
Excep.	20	5	0.34
Excep.	17	6	0.35
Excep.	14	7	0.33
Regular	13	8	0.35
Regular	11	9	0.34
Regular	10	10	0.34
Total	293		T_{NoRule} 3.42
With a Rule			
Excep.	Freq.	rank	Time Complexity
Excep.	100	1	$0.24 = 100/293 \times 1 \times 0.7$
Excep.	50	2	$0.24 = 50/293 \times 2 \times 0.7$
Excep.	33	3	$0.24 = 33/293 \times 3 \times 0.7$
Excep.	25	4	$0.24 = 25/293 \times 4 \times 0.7$
Excep.	20	5	$0.24 = 20/293 \times 5 \times 0.7$
Excep.	17	6	$0.24 = 17/293 \times 6 \times 0.7$
Excep.	14	7	$0.23 = 14/293 \times 7 \times 0.7$
Total			1.67
Regular			Time
Regular	13		
Regular	11		
Regular	10		
Total			2.1 = 7×0.3
T_{Rule}			3.77 = $1.67 + 2.1 > 3.42$

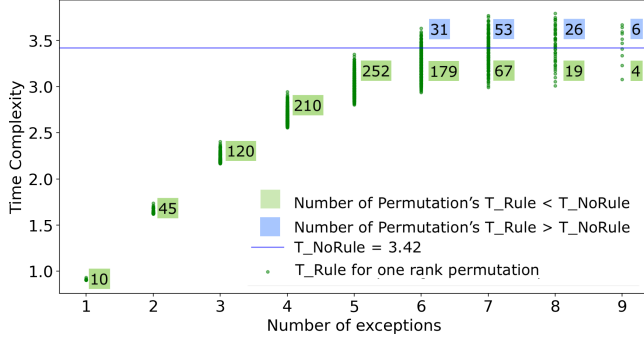


Figure 2: Time complexity of all rank permutations with 1-9 exceptions

In order to comprehensively compare T_{Rule} and T_{NoRule} with different rank permutations, we created data sets with 10 items having 1 to 9 exceptions. For each data set, we exhaustively tested all the rank permutations for the exceptions and regulars by calculating T_{Rule} for each permutation. As shown in Figure 2, time complexity varies depending on the rank permutation, rather than being constant. When the regulars are in the highest ranks, time complexity is smallest. When the regulars are in the lowest ranks, time complexity is largest.

Further, given some permutations, the minimum T_{Rule} for all numbers of exceptions is smaller than T_{NoRule} , yielding a rule no matter how many exceptions there are. The number of permutations where $T_{Rule} < T_{NoRule}$ are labeled in Figure 2. Even in an extreme scenario with nine exceptions and one regular item, where it is impossible to derive a rule, following the TP will provide a rule in four of the possible permutations where $T_{Rule} < T_{NoRule}$.

We tested two larger N conditions (100, 1000) to determine whether rank would continue to play a role. The item frequencies in both data sets follow a Zipfian distribution, with the lowest frequency set at 10. Given the large N , it is impossible to exhaustively calculate the T_{Rule} for all permutations³. Instead, we calculated the minimum T_{Rule} where the regulars are all the top ranked items and the maximum T_{Rule} where the regulars are all the bottom ranked items⁴. Figure 3 shows the $T_{Rule(MIN)}$ and $T_{Rule(MAX)}$ for $N = 100, 1000$ with different numbers of exceptions.

The time complexity shows a quadratic pattern and intersects with T_{NoRule} at two points: at the rising part of the function in purple ($e = \theta_{min}$) and at the end of the function in yellow ($e = \theta_{max}$), suggesting that a rule will be derived ($T_{Rule} \leq T_{NoRule}$) when $e \leq \theta_{min}$ or $e \geq \theta_{max}$. The summary of θ_{min} and θ_{max} is listed in Table 5.

For $T_{Rule(MAX)}$, $\theta_{max} > N$, therefore only θ_{min} is valid.

³When there are 100 items with 20 regulars, there are 5.36×10^{20} rank permutations.

⁴When there are 100 items and 20 regulars, $T_{Rule(MIN)}$ is derived when 20 regulars are ranked from 1-20; $T_{Rule(MAX)}$ is derived when the regulars are ranked from 81-100.

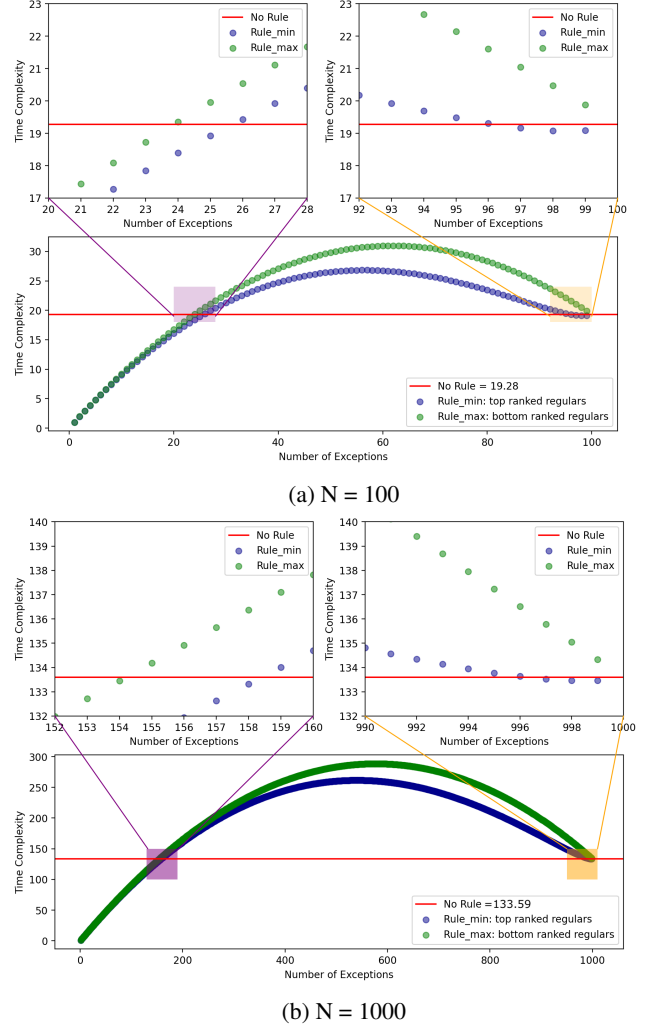


Figure 3: The maximum and minimum T_{Rule} plot for $N = 100, 1000$ with different number of exceptions

Table 5: The θ_{min} and θ_{max} for $T_{Rule(MIN)}$ and $T_{Rule(MAX)}$

	$N = 100$	$N = 1000$
T_{NoRule}	19.28	133.59
$\theta = N/\ln(N)$	21.71	144.76
$T_{Rule(MAX)}$ (with integer θ)		
θ_{min}	23	154
θ_{max}	NA(> 100)	NA(> 1000)
$T_{Rule(MIN)}$ (with integer θ)		
θ_{min}	25	158
θ_{max}	97	997

A rule is derived when $e \leq \theta_{min}$ or $e \geq \theta_{max}$

However, for $T_{Rule(MIN)}$, $\theta_{max} < N$: $\theta_{max} = 97$ when $N = 100$; $\theta_{max} = 997$ when $N = 1000$. This result suggests that when there are only 1, 2, or 3 regulars in 100 and 1000 items, a rule can still be derived since $T_{Rule} < T_{NoRule}$, which is implausible. In addition, the θ_{min} is different for $T_{Rule(MAX)}$ and

Table 6: Summary of the corpus data for each child.

	Age range	files	Total Verb Types (N)	Irregular Verb Types (e)	Total Verb tokens	Irregular tokens	Corpus
Adam	2;3 - 2;11	18	306	62	6,747	3,632	Brown (1973)
Eve ⁵	1;6 - 1;8	5	93	36	564	337	Brown (1973)
Sarah	2;3 - 2;10	33	189	48	1,759	1,035	Brown (1973)
Peter	1;3 - 2;6	14	424	67	7,532	3,647	Bloom (1973)
Naomi	1;3 - 1;11	20	128	43	1,240	757	Sachs (1983)
Allison	1;5 - 2;11	6	88	36	612	335	Bloom, Hood, and Lightbown (1974)
April	1;10 - 2;1	2	50	19	128	62	Higginson (1985)
Fraser	2;0 - 2;5	90	371	78	13,924	9,903	Lieven, Salomo, and Tomasello (2009)

$T_{Rule(MIN)}$. When $N = 100$, if the regulars are ranked at the top of the frequency list, a rule could be derived when there are fewer than 25 exceptions; if the regulars are ranked at the bottom, there can be fewer than 23 items. When there are 24 and 25 exceptions, a rule might be derived given the proper permutation.

Similarly, when $N = 1000$, if the regulars are all the most frequent items, a rule can be derived when there are fewer than 158 exceptions; if the regulars are all the least frequent items, a rule can be derived when there are fewer than 154 exceptions. When there are 155-158 exceptions, some permutations allow a rule to be derived while others do not.

Summary

In summary, we have shown that the frequency rank of the regulars substantially affects the time complexity, thus influencing the TP's prediction. Data sets with the same number of exceptions but different rank permutations lead to inconsistent results. We conclude that using $f_i \cdot r_i = C$ in a Zipfian distribution to eliminate rank as a variable in the TP's derivation for T_{Rule} is not appropriate.

Because the T_{Rule} is quadratic, two sets of data will fit the rule-deriving criterion $T_{Rule} < T_{NoRule}$: $e \leq \theta_{min}$ or $e \geq \theta_{max}$. That result casts doubt on whether the cognitive motivation for the TP can be maintained. The cognitive motivation is that a rule reduces lexical access time by minimizing the access time of the regulars. The more the regulars (or the fewer the exceptions), the less lexical access time. That assumes that lexical access time is linearly related to the number of exceptions, but in our simulations, the relationship between lexical access time and the number of exceptions is quadratic. This leads to some unlikely predictions, such as derivation of a rule when there are 97 exceptions in 100 items and 997 exceptions in 1000.

Testing the TP on Corpus Data

Our hypothetical data have demonstrated inconsistencies in the TP and unlikely predictions. Here we test corpus data,

⁵Eve's data are different from Yang's count because Yang made an error. Yang counted Eve's data from 1;6 to 1;10. However, Eve made the first overregularization error at the age of 1;8 (Brown/Eve/010800.cha), when she said 'I *seed it'.

since many of our hypothetical scenarios may never exist.

Yang (2016)'s Tests

Yang (2016) applied the TP to Adam's and Eve's data from the Brown corpus (Brown, 1973). The first instance of an over-regularization in a child's longitudinal corpus can be seen as an unambiguous marker for the presence of a productive '-ed' rule for the past tense. The child may have formulated the rule earlier, but has definitely formulated it by the time of the first over-regularization. For both Adam and Eve, there were more exceptions than predicted before the first over-regularization error, discrepancies that Yang attributed to sampling error.

New Tests

We replicated Yang's method on eight children's longitudinal data, including Adam and Eve, from CHILDES (MacWhinney, 2000). We tabulated the number of irregular verbs (e). We computed several values: the approximated $\theta = N/\ln N$, the numerical θ (by solving inequation (1)), T_{Rule} and T_{NoRule} via the verbs' probability and rank. The eight children's past tense acquisition has been extensively studied in the previous literature⁶, and their data are shown in Table 6.

Adopting Yang's method, we included all the files from the first recording to the file where the child made her/his first over-regularization error and counted all verb forms. The sample age and density vary across the children. The average age range is about 8 months, with a minimum of 2 months (Eve) and maximum of 18 months (Allison). The average file number is 23.5, with a minimum of 2 files (April) and maximum of 90 files (Fraser). All the verbs were first automatically extracted from the annotated corpora in CHILDES using the NLTK python package (Bird, Klein, & Loper, 2009), and were hand-checked by a human annotator.

We compare the number of irregulars (e) with the approximated $\theta_{TP} = N/\ln N$ and the numerical θ_n , as shown in Table 7. θ_{TP} and θ_n are confirmed only for Peter, whose irregular verbs ($e = 67$) are fewer than the TP's approximated $\theta_a = 70.1$

⁶Adam, Eve, Sarah, Peter, Naomi, Allison and April were studied in Marcus et al. (1992). Fraser was studied in Lieven et al. (2009).

Table 7: Results of comparing e vs θ and T_{Rule} vs T_{NoRule}

	N	e	θ_{TP}	$e < \theta_n$	θ_n	$e < \theta_n$	T_{NoRule}	T_{Rule}	$T_{Rule} < T_{NoRule}$
Adam	306	62	53.5	×	57.0	×	33.80	51.33	×
Eve	93	36	20.5	×	22.9	×	17.51	25.11	×
Sarah	189	48	36.1	×	38.1	×	25.65	37.81	×
Peter	424	67	70.1	✓	74.8	✓	43.82	57.74	×
Naomi	128	43	26.4	×	28.2	×	19.63	31.23	×
Allison	88	36	19.7	×	21.0	×	18.24	34.72	×
April	50	19	12.8	×	13.7	×	14.64	14.29	✓
Fraser	371	78	62.7	×	66.5	×	26.34	60.03	×

$\theta_{TP} = N/\ln N$ is the approximated θ . θ_n is the numerical θ calculated by solving $\theta_n/N(\theta_n/H_{\theta_n} - \theta_n) + \theta_n = N/H_N$

and the Harmonic N-based $\theta_n = 74.8$. The other seven children’s data do not confirm the TP’s prediction that $e \leq \theta$.

We also use each child’s data to compare T_{NoRule} and T_{Rule} with each verb’s probability and rank. That is, we rank all the verbs according to their frequencies and calculate the time complexity based on their probabilities and ranks. The results are summarized in Table 7. The TP predicts that, for all the children, $T_{Rule} < T_{NoRule}$, since they have already applied the rule. That holds only for April. For the other seven children, T_{Rule} is larger than T_{NoRule} , suggesting that no rule should be derived for these children.

In summary, most children’s data do not conform to the TP’s prediction. For the eight children we tested, only Peter’s actual number of exceptions is smaller than the threshold and only for April is $T_{Rule} < T_{NoRule}$. The data are not definitive, for at least three reasons. The children may have already acquired the rule before making over-regularization errors. Children’s first errors may have occurred outside of taping sessions. The sample of children’s speech might be a small fraction of children’s vocabulary.

Discussion

Our tests of the TP on hypothetical data and corpus data suggest that the original the TP makes inconsistent predictions on hypothesized data and is not confirmed by empirical corpus data. We have shown that the approximated $\theta = N/\ln(N)$ yields a smaller number of permissible exceptions than the numerical θ ranging from 10 to 1000. We have also shown that the approximated and simplified result of inequation (1) ($e \leq N/\ln N$) changes a quadratic function of e to a linear function; the quadratic function for criterion $T_{Rule} < T_{NoRule}$ leads to two different θ thresholds instead of a single threshold.

Our main theoretical contribution is to show much rank permutation affects the T_{Rule} . For example, with ten items, depending on the rank permutation, $T_{Rule} < T_{NoRule}$ even if there are nine exceptions. Our main empirical contribution is to show that only one child (Peter) of the eight we examined has fewer exceptions than those predicted by both forms of the model, and for only one child (April) is $T_{Rule} < T_{NoRule}$.

These discrepancies could stem from the inconsistency between the TP’s cognitive motivation (a rule is derived to save

time by compressing the lexical access time of the regulars) and its calculation of time complexity in inequation (2). The cognitive motivation presumes that the time complexity with a rule has a linear relationship with the number of exceptions, that T_{Rule} increases as e increases. However, the formula of the T_{Rule} has a quadratic relationship with e . To retain more of the original model and resolve the inconsistencies we have described, one can either modify the time complexity calculation or the cognitive motivation underpinning its use. If we want to keep the idea that a rule is derived to save time, and use the number of exceptions to predict when a rule is deployed, time complexity should be at least a monotonically increasing function of the number of exceptions. The current time complexity is calculated based on the Zipfian distribution and serial search hypothesis, which yields a quadratic relationship with the number of exceptions and is affected by the rank permutation. Alternately, the time complexity can be calculated based on a uniform distribution, which would get rid of the rank influence; or based on another retrieval model, instead of serial search, that could produce a linear relationship between time complexity and the number of exceptions. On the other hand, the cognitive motivation could be modified to incorporate other motivations for a rule to be derived, such as optimizing both time complexity and memory space. Although the current version of the TP is not adequate enough to explain the rule-deriving process, it provides some insight into how we can approach the rule-deriving process computationally. With some theoretical and mathematical modifications, the TP can still be a plausible hypothesis to explain when a rule is used.

References

- Anderson, S. R. (1969). *West scandinavian vowel systems and the ordering of phonological rules*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.
- Berko, J. (1958). The child’s learning of English morphology. *Word*, 14(2-3), 150–177.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc.
- Bloom, L. (1973). *One word at a time: The use of single word utterances before syntax* (Vol. 154). Walter de Gruyter.

- Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive psychology*, 6(3), 380–420.
- Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.
- Emond, E., & Shi, R. (2021). Infants' rule generalization is governed by the tolerance principle. In *Proceedings of the 45th Annual Boston University Conference on Language Development* (p. 191-204). Somerville, MA.
- Forster, K. I. (1976). Accessing the mental lexicon. *New approaches to language mechanisms*, 257–287.
- Forster, K. I. (1992). Memory-addressing mechanisms and lexical access. *Orthography, phonology, morphology, and meaning*, 413.
- Halle, M., & Marantz, A. (1993). *Distributed morphology and the pieces of inflection., the view from building 20*. MIT Press, Cambridge, Mass.
- Higginson, R. P. (1985). *Fixing: Assimilation in language acquisition*. Unpublished doctoral dissertation, Washington State University.
- Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3), 481–507.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.
- Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, i–178.
- Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, 111(3), 721.
- Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11), 456–463.
- Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's Language*, 4, 1–28.
- Schuler, K., Yang, C., & Newport, E. (2021). Testing the tolerance principle: Children form productive rules when it is more computationally efficient. *PsyArXiv*.
- Yang, C. (2005). On productivity. *Linguistic Variation yearbook*, 5(1), 265–302.
- Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.
- Yang, C. (2018). A user's guide to the tolerance principle. *Manuscript. University of Pennsylvania (ling. auf. net/lingbuzz/004146)*.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley Press.
- guments. We thank Baskaran Sripathmanathan for helpful conversations on the math part.

Acknowledgement

We thank Charles Yang for the helpful discussion that he generously provided despite his disagreements with our ar-