Testing the Tolerance Principle on Corpus Data

1 Introduction

1.1 Deriving the Tolerance Principle

Do children learn a rule for forming the past tense? Or do they simply have a list of verbs and local regularities? If they do have a rule, how is it formed, given that the input is noisy? (Yang, 2016) has proposed the Tolerance Principle (TP) to explain how children deploy general (productive) rules given noisy input. We first describe the principle then our revision of it. Yang assumes that the humans apply the Elsewhere Condition (e.g. McClelland and Rumelhart, 1981; Plaut, 1997) in processing rules and exceptions. The Elsewhere Condition can be implemented as a serial search procedure in which each lexical item is compared to all the exceptions to the general rule. If a match is found, a specific rule for the matching exception is triggered. If not, the general rule is applied as shown in (1):

(1) IF w = e_1 THEN apply rule e_1 to w...
IF w = e_2 THEN apply rule e_2 to w...
IF w = e_3 THEN apply rule e_3 to w...
...
IF w = e_N THEN apply rule e_N to w...
IF w not in set{e_1,e_2,e_3,e_4....e_N}, THEN apply general rule to w...

For example, to retrieve the past tense form for the verb *eat*, *eat* is first compared to the exceptions in the irregular verb inventory. When it is found, the irregular rule for *eat* is applied to derive the past tense form *ate*. To retrieve the past tense form for the verb *type*, when the search for a match in the irregular verb set fails, the general past tense rule is applied to derive *typed*.

Since lexical retrieval is a self-terminating serial search process, rules and exceptions are organized to minimize the time required. A productive rule to produce a set of N items is derived when applying the rule takes less time than does processing all the items individually. Without a productive rule, all the items are ranked by their frequencies, so that frequently used items will be processed most quickly. When a productive rule is generated, items are separated into two categories, regulars and exceptions. The exceptions are ranked by frequency and the rule is applied only when the item is not found in the set of exceptions. For example, when there is no productive rule and the vocabulary inventory has *n* regular items (*w*) and *m* irregular items (*e*), all the items are arranged according to their frequency, as shown in (2). When there is a productive rule, all the irregular items are ranked based on frequency, and all the regular items are concatenated into a set where one rule will be applied, as shown in (3).

(2) Without productive rule:

$$N = n+m \begin{cases} w_1 & frequency: 100 \\ e_1 & frequency: 99 \\ w_2 & frequency: 98 \\ ... & frequency: ... \\ w_m & frequency: 2 \\ e_n & frequency: 1 \end{cases}$$

(3) With productive rule:

$$N = n+1 \begin{cases} e_1 & frequency: 99 \\ e_2 & frequency: 90 \\ e_3 & frequency: 87 \\ ... & frequency: ... \\ e_n & frequency: 1 \\ & w_1, w_2, w_3 ... w_n \end{cases}$$

The rule will be generated only when the cost (Yang refered to the cost as 'time complexity') to process the list with the rule is less than processing all the items. As shown in (2) and (3), there are fewer items (n+1) to process in the list with a productive rule than in the list without a productive rule (n+m). On occasion, however, the time to search for an item in the list *with* a rule can be more than the search time in the list *without* a rule. For example, for the item w_2, which is a more frequent regular word, it will take less time to find w_2 in the list without a productive rule than in the list with one. Since w_2 is the third most frequent word, the search only needs to examine w_1 and e_1 before finding a match for w_2. In the list with a productive rule, the regular item will be reached only after all the exceptions are checked, which creates more cost to find w_2. Therefore, the rule will be deployed only when the average cost to search for each word in the list with a productive rule (T_R) is less than the list without a rule (T_N).

(4) $T_N > T_R$

According to Yang, the average cost to search for a word is the product of the probability of the word and the retrieval time for the word. To calculate the probability for each word, Yang assumes that any sample from a large corpus follows the Zipfian distribution; therefore, the product of the frequency of the word (f_i) and the rank (r_i) of the word is a constant (C).

(5) $r_i f_i = C_i$

The probability of a word in a corpus (p_i) is the frequency of the word (f_i) divided by the sum of the frequencies of all the words. The probability of occurrence (p_i) for w_i can be expressed as (6) where $H_N = \sum_{k=1}^{N} \frac{1}{r_k}$.

(6) $$p_i = \frac{f_i}{\sum_{k=1}^{N} f_k} = \frac{\frac{C_i}{r_i}}{\sum_{k=1}^{N} \frac{C_k}{r_k}} = \frac{\frac{1}{r_i}}{\sum_{k=1}^{N} \frac{1}{r_k}} = \frac{1}{r_i \cdot H_N}$$

Since all the items are stored in a list ranked according to frequency, the retrieval time for each item is determined by the rank of the item. For example, the word *the* appears more frequently in the corpus than the word *give*, and therefore it costs less to retrieve. Yang simplified this as 'the *i*-th ranked item takes *i* units of time to be retrieved' (Yang, 2018b). The cost for w_i (T_{w_i}) is shown in (7). The total cost for N items in the list (T_N) is shown in (8).

(7) $$T_{w_i} = r_i \cdot \frac{1}{r_i \cdot H_N} = \frac{1}{H_N}$$

(8) $$T_N = \sum_{k=i}^{N} (r_i \cdot \frac{1}{r_i \cdot H_N}) = \frac{N}{H_N}$$

If a rule is used, all the exceptions are stored in a ranked list and all the regular items are stored in a set after the list of exceptions. The exception list is processed in the same way as is the list without rules. If there are e items in the exceptions, the cost for e_i (T_{e_i}) in the exception list is shown in (9). The total cost for all the items in the exception list (T_e) is shown in (10).

(9) $$T_{e_i} = r_i \cdot \frac{1}{r_i \cdot H_e} \cdot \frac{e}{N} = \frac{1}{N \cdot H_e}$$

(10) $$T_e = \sum_{k=1}^{e} (r_i \cdot \frac{1}{r_i \cdot H_e} \cdot \frac{e}{N}) = \frac{e}{H_e} \cdot \frac{e}{N}$$

The cost for all regulars is the same, which is the constant e, given that there are e items in the exception list. The total cost to process all the regular items ($T_{\overline{w}}$) is:

(11) $$T_{\overline{w}} = (1 - \frac{e}{N}) \cdot e$$

The total cost for a list with a productive rule (T_R) is the sum of the cost for exceptions and regular items, as shown in (12):

(12) $$T_R = T_e + T_{\overline{w}} = \frac{e}{N} \cdot \frac{e}{H_e} + (1 - \frac{e}{N}) \cdot e$$

A productive rule will be derived only when T_R is smaller than T_N. To derive the maximum number for the exceptions (e), first we approximate the Nth harmonic number with the natural log (ln$N$), then we make $T_R \leq T_N$:

(13) $$\begin{aligned} T_R &\leq T_N \\ \frac{e}{N} \cdot \frac{e}{H_e} + (1 - \frac{e}{N}) \cdot e &\leq \frac{N}{H_N} \\ \frac{e}{N} \cdot \frac{e}{lne} + (1 - \frac{e}{N}) \cdot e &\leq \frac{N}{lnN} \\ \frac{e^2}{N} \cdot (\frac{1}{lne} - 1) + e &\leq \frac{N}{lnN} \end{aligned}$$

Since $\frac{e^2}{N} \cdot (\frac{1}{lne} - 1)$ is always smaller than or equal to zero, as long as e is smaller than $\frac{N}{lnN}$, then T_R is always smaller than T_N. Thus, the TP is derived:

(14) *Tolerance Principle*
Let R be a rule applicable to N items, of which e are exceptions. R is productive if and only iff:
$e \leq \theta_N, where \theta_N = \frac{N}{lnN}$ (Yang, 2016, p.64)

1.2 Why should(n't) the Tolerance Principle (TP) work?

Recall that the TP assumes the Elsewhere Condition, and is derived based on the assumption that lexical retrieval is a serial search of a frequency-ranked list that results in the logarithmic relationship between frequency and retrieval time (Murray and Forster, 2004). There is no guarantee that it corresponds to the actual psychological process learning. The TP appeaingly handles a critical fact about language acquisition: children generate rules based on noisy input. The TP provides an elegant and succinct way to quantify the 'noise' in the input.

Evidence from artificial language learning (Schuler et al., 2016) supports the TP. Children between the ages of 5 and 7 heard names of nine novel objects in both singular and plural forms. Each plural marker either followed a rule (add *ka*) or instead used an individual suffix (add *po, tay, lee bae, muy,* or *woo*). In one condition, children heard five nouns with the *ka* marker and four with individual markers. In another condition, they heard three nouns with the *ka* marker and six with individual markers. As the TP predicts, children learned the rule under the 5/4 condition but not the 3/6 condition, as shown by their ability to use *ka* as a general plural marker in a Wug-like test.

Despite this evidence, other research has queried whether the TP can be applied to explain language acquisition in real life.

Yang used lnN as the approximation of the Harmonic number, as shown in (13). For smaller N, however, $\log_2 N$ is a better approximation of H_N (Andreae and Kuiper, 2018). For example, in Schuler et al. (2016)'s experiment, when N = 9, $H_9 \approx 2.83$, while $\ln 9 \approx 2.20$ and $\log_2 9 \approx 2.71$. For N = 9, the base-2 logarithm is a better approximation than the natural logarithm. That the values matter can be seen by looking at the 4 exceptions, which were calculated based on $\theta = N/lnN$, where $9/2.20 \approx 4.10$. If, instead, the original $H_9 \approx 2.83$ were inserted, then $\theta = N/H_N$ would be $9/2.83 \approx 3.18$. That result produces a tolerance threshold of three exceptions, which goes against the results in Schuler et al. (2016).

The TP also received criticism on the more general level of how to approach acquisition. For example, perhaps communicative needs, context, prior learning and cognitive load should be taken into account (Goldberg, 2018). As another example, serial search might be a flawed model compared to a more relevant lexical retrieval model (Kapatsinski, 2018)[1].

In addition, Yang examined Adam's and Eve's corpus data on past tense verbs to test the TP. However, the results didn't conform to TP predictions. There are more irregular verbs in Adam's and Eve's data than the maximum number of exceptions (θ) the TP predicts. Yang attributed the discrepancy to sampling effects.

In this paper, we developed a method to test the TP on corpus data and explored how would sampling effects influence the applicability of the TP on the corpus data. The rest of the paper is organized as follow: in section 2, we revisited Yang's test on Adam and Eve and proposed a revised methods that would be more appropriate for corpus data testing; in section 3, we tested the revised method on eight children's corpus data, including Adam and Eve; in section 4, we used Fraser's corpus (a densely sampled corpus from Manchester corpus) to explore the sampling effects on the TP.

2 Revised Testing Methods

2.1 Yang's Test on Adam's and Eve's Data

In Chapter 4 of Yang (2016)'s book, he established a procedure for the applying the TP. In our study, we followed his procedure.

(15)
  a. Obtain a rule R along with its structural description and structural change.
  b. Count N, the number of lexical items that meet the structural description of R.
  c. Count e, the subset of N that are exceptions to R.
  d. Compare e and the critical threshold $\theta_N = \frac{N}{lnN}$ to determine productivity.

Yang applied this procedure to explain the acquisition of past tense in English children. English speaking children usually start to produce the past-tense form by the age of 2. Most children also produce overregularization errors on past tense, such as *grewed*, *feeled* (e.g. Marcus et al., 1992). The first instance of an overregularization error can be seen as an unambiguous marker for the presence of a productive 'add *-d*' rule for past tense.

Adam produced his first overregularization error at the age of 2;11, when he said *What dat feeled like?* (Brown, 1973). This error implied that Adam had already constructed the past tense rule. According to the TP's prediction, the number of irregular verbs that Adam knew (e) must be smaller than $\theta = N/lnN$, where N is the number of all the verbs in his vocabulary. Adam's first recording starts at 2;3. Yang thus estimated Adam's effective vocabulary (N) as all the verbs he produced between 2;3 and 2;11. Yang did not only count all the past

[1] See Yang (2018a) for his response to Goldberg and Kapatsinski.

tense verbs, he counted all forms of verbs as N. According to Yang, as long as Adam produced one form of a verb, that verb has to be in Adam's lexicon. Based on this method, he found 300 verbs, which made N = 300. Therefore, $\theta = N/lnN \approx 53$, which means Adam can learn the rule when there are fewer than 53 irregular verbs. However, Yang counted 57 irregular verbs in Adam's total 300 verb lexicon. He attributed the difference between 57 and 53 to sampling effects.

Yang used the same method to test Eve's data. Eve's first overregularization error appeared at 1;10 when she said *it falled in the briefcase*[2] (Brown, 1973). Yang found 163 verbs Eve produced between the age 1;6, when Eve had her first recording, and 1;10. When $N = 163$, $\theta = N/lnN \approx 32$, which means Eve could only tolerate 32 irregular verbs in order to produce the past tense rule. However, Yang found 49 irregulars in her production, which is again higher than what the TP predicts. He attributed the difference to undersampling of Eve's data.

2.2 Revised Testing Methodology

In Yang's test, the TP failed to account for Adam's and Eve's corpus data on past tense acquisition. With the proposed new methodology, we aim to preserve Yang's insight of the TP, which is that a rule will be derived if the cost to retrieve an item from a list with a rule is smaller than from a list without a rule. We develop a different version of the TP and altered the formula to calculate the cost to retrieve an item.

First, we aim to better estimate the probability of each item (p_i) in a list or items ranked by frequency. Yang assumed a Zipfian distribution for all items N and all the exceptions e. However, a Zipfian distribution is not guaranteed for a small corpus, such as all the verbs and irregular verbs a 2-year-old child knows, which affects how p_i is derived. Yang's formula for p_i (16b) is based on a convenient fact of Zipfian distribution: when a corpus follows a Zipfian distribution, the product of the frequency of a word (f_i) and the rank of that word (r_i) is a constant C,shown in (16a). This is derived from the formal expression of Zifp's law, which is shown in (17): the frequency of the rth most frequent word is inversely proportional to its rank, where the exponent (α) equals 1. Formula (16a) is only valid when α is 1, when N has a Zifpian distribution. When a smaller corpus (such as children's effective vocabulary of verbs and irregular verbs) has a power law distribution but doesn't necessarily a Zifpian distribution, where the exponent (α) is not 1, formula (5) is no longer valid, thus p_i needs to be recalculated.

(16) a. $r_i \cdot f_i = C$ (replicate of (5))

b. $$p_i = \frac{f_i}{\sum_{k=1}^{N} f_k} = \frac{\frac{C_i}{r_i}}{\sum_{k=1}^{N} \frac{C_k}{r_k}} = \frac{\frac{1}{r_i}}{\sum_{k=1}^{N} \frac{1}{r_k}} = \frac{1}{r_i \cdot H_N} \quad \text{(replicate of (6)}$$

(17) $f_i = Cr_i^{-\alpha}, \alpha = 1$

In this paper, we propose to measure the actual corpus distributions of all the verbs (N) and the irregular verbs (e), and use the empirically estimated exponents that best fit those ranked frequency distributions in our calculations. Since all the verbs and all the irregular verbs do not necessarily share the same distribution, we will use α and β to represent the exponents for these two distributions respectively. To include the exponent as a variable, the probability formula can be written as follow, where $H_{n,m} = \sum_{k=1}^{n} \frac{1}{k^m}$:

(18) Probability of occurrence for ith ranked word (p_i):

$$p_i = \frac{\frac{1}{r_i^{\alpha}}}{\sum_{k=1}^{N} (\frac{1}{r_k^{\alpha}})} = \frac{1}{r_i^{\alpha} \cdot H_{N,\alpha}}$$

Based on the new formula for p_i, the cost to retrieve an item from a list without rules (T_N) and the cost to retrieve an item from a list with rules (T_e) can be written as follow:

(19) Cost for a list of N items without a productive rule (T_N):

$$\begin{aligned} T_N &= \sum_{k_1}^{N} (r_i \cdot p_i) \\ &= \sum_{k=1}^{N} (r_i \cdot \frac{1}{r_i^{\alpha} \cdot H_{N,\alpha}}) \\ &= \frac{H_{N,\alpha-1}}{H_{N,\alpha}} \end{aligned}$$

[2]Yang made an error here. Eve made the first overregularization error at the age of 1;8, when she said *I seed it*.

(20) Cost for a list with e exceptions and a productive rule (T_R):

$$T_R = \sum_{k_1}^{e}(r_i \cdot p_i) \cdot \frac{e}{N} + (1 - \frac{e}{N}) \cdot e$$
$$= \frac{H_{e,\beta-1}}{H_{e,\beta}} \cdot \frac{e}{N} + (1 - \frac{e}{N}) \cdot e$$

(21) A productive rule will be derived when $T_R \leq T_N$:

$$\frac{H_{e,\beta-1}}{H_{e,\beta}} \cdot \frac{e}{N} + (1 - \frac{e}{N}) \cdot e \quad \leq \frac{H_{N,\alpha-1}}{H_{N,\alpha}}$$

Unlike formula (13) where H_N can be conveniently approximated using lnN, there is no mathematical approximation for the Harmonic number in the inequation (21). Therefore, the new version of the TP is not going to produce a maximum number of the irregular items; instead, we propose to compare T_R and T_N directly. The Tolerance Principle will be confirmed if T_R is smaller than T_N as predicted in (21).

In the next section, we extracted all the variables (e, N, α and β) from eight children's corpora to compare T_N and T_R. Instead of counting all the verb types the child produced as the N, we also estimate the N by using the verb types from parents input. Since N represents the child's effective vocabulary, children's production and parents' input represent the lower and upper boundary of the vocabulary respectively. The type of irregular verbs e is counted as long as one form of the irregular verb (not necessarily the past tense form) appear in children's production or parents' input. The distribution of N and e are then mapped to the best fitted power law function to calculate the exponent α and β.

3 Testing the TP on corpus data

3.1 Testing on Adam's, Eve's and six other children's corpus data

In this section, we use the revised testing methods to test eight children's corpus data on their past tense acquisition. The age of the first recording and the age of the first overregularization error for each child is shown in TABLE 1, with a summary of the each child's corpus data.

Table 1: Summary of corpus data for each child

	Age of first recording to first overregularizaiton error	corpus	files	words	input words	verbs	input words
Adam	2;3 - 2;11 (*feeled*)	Brown (1973)	18	39403	30366	6747	4670
Eve	1;6 - 1;8 (*seed*)		5	5304	11253	564	1618
Sarah	2;3 - 2;10 (*topped*)		33	18778	27682	1759	3867
Peter	1;3 - 2;6 (*broked*)	Bloom et al. (1974)	14	52769	95180	7532	15537
Naomi	1;3 - 1;11 (*doed*)	Sachs (1983)	20	8009	9634	1240	1463
Allison	1;5 - 2;11 (*throwed*)	Bloom (1973)	6	4605	9366	612	1453
April	1;10 - 2;1 (*boughted*)	Higginson (1985)	2	1376	4435	128	658
Fraser	2;0 - 2;5 (*seed*)	Lieven et al. (2009)	90	137407	222200	13924	32359

All of the data were automatically extracted from the annotated corpora in CHILDES using the NLTK python package. The verbs in each file were identified using part-of-speech taggers annotated by the MOR program (MacWhinney, 2012). The number of verb types and irregular verb types in parents' input (U_p and e_p) and in children's production (U_c and e_c) are shown in TABLE 2, with the exponents for the verb types of parents' input (α_p), children's production (α_c) and the exponents for the irregular verb types in parents' input (β_p) and the children's production (β_c). The log-log graphs for each child can be found in the Appendix.

We used U_c and U_p to represent N separately and inserted the value of the variables (e, α, β) to formula (19) and (20) to calculate T_N and T_R. The results for T_R and T_N for each child is shown in TABLE 3.

As shown in TABLE 3, the TP successfully predicted five children's acquisition on past tense acquisition, that the past tense rule is derived because T_R is smaller than T_N. However, three children's data (Eve, Allison and April) do not support the TP's prediction. This could be attributed to the effect of smaller sample size, since Eve, Allison and April have less data between the first recording and the first appearance of the overregularization error than other children (as shown in TABLE 1). In order to further test how would sample size affect the testablibity of the TP, we used Fraser's corpus to explore the sampling effects on the TP.

Table 2: Number of observed total number of verb types, irregular verbs and exponents

	U_p	α_p	U_c	α_c	e_p	β_p	e_c	β_c
Adam	275	0.69	270	0.66	70	0.64	62	0.61
Eve	136	0.74	91	0.84	50	0.65	36	0.73
Sarah	293	0.71	189	0.77	68	0.58	48	0.62
Peter	633	0.64	424	0.69	83	0.51	67	0.54
Naomi	174	0.81	128	0.76	59	0.66	43	0.66
Allison	140	0.77	88	0.87	44	0.65	36	0.84
April	100	0.84	50	1.23	37	0.80	19	1.23
Fraser	581	0.56	371	0.60	97	0.44	78	0.49

Table 3: Comparison between observed cost and the TP predicted cost

	N = Children's production (U_c)			N = Parent's input (U_p)		
	T_R	T_N	$T_R \leq T_N$	T_R	T_N	$T_R \leq T_N$
Adam	52.53	78.07	True	57.92	76.24	True
Eve	26.27	22.55	**False**	37.66	36.94	**False**
Sarah	39.94	47.90	True	57.62	78.67	True
Peter	60.18	115.04	True	76.05	181.99	True
Naomi	32.94	34.03	True	50.37	55.54	True
Allison	25.46	21.03	**False**	34.65	36.42	True
April	13.44	8.13	**False**	27.34	24.48	**False**
Fraser	67.27	110.64	True	86.71	181.58	True

3.2 Exploring Small Sample effects on the TP

Fraser is the most densely sampled corpus in this study. His first recording and first overregularization error was 5 month apart and he had 90 recording files. He was recorded for five hours per week in the first month (2;0 to 2;1) and one hour per week for the rest of the four months (2;2 to 2;5). The densely sampled corpus captured 358 types of verbs and 78 types of irregular verbs that Fraser produced, and 566 types of verbs and 97 types of irregular verbs in parents' input. Eve's, April's and Allison's corpus are less densely sampled comparing to Fraser's. Eve's first recording and first overregularization error was only 2 month apart and she only had 5 recording files. She was recorded twice a month between the age of 1;6 - 1;8. April's first recording and first overregularization error was 3 months apart and she was only recorded twice (1;10 and 2;1). Allison's first recording and first overregularizaiton error was 18 months apart and she was recorded only six times. Four of her 6 files were recorded before 2;0 (1;5, 1;7, 1;8 and 1;10) and only 2 files were recorded after she turned two years old (2;4 and 2;10). The short intervals between the age of overregularization error and the first recording (such as Eve and April) could be the reason for a smaller sample since the children simply didn't produce enough verbs in such short time intervals. Or, the smaller sample could be a result of sparse sampled data (such as Alison) that the recordings failed to cover the children's longitudinal development. In this section, we used Fraser's corpus to explore these two types of small sample.

We first investigated the age related small sample size effect by setting the age of first recording as 2;4, only one month before Fraser made the first overregularization error (2;5). There are 11 recording files between 2;4 - 2;5 in Fraser's corpus, and a summary of the corpus data is shown in TABLE 4. Then, we randomly selected 3,4,5,6 files from Fraser's corpus to represent different density of the corpus from age 2;0 - 2;5. The summary of the corpus is also shown in TABLE 4. Subsequently, the number of verb types and irregular verb types in children's production (U_c and e_c) and parents' input (U_p and e_p), and the exponent for the distribution of all verbs (α_c and α_p) and irregular verbs (β_c and β_p) are subtracted from the data, shown in TABLE 5. These variables were then inserted into formula (19) and (20) to calculate the T_N and T_R. The results are shown in TABLE 6.

As shown in TABLE 6, the TP was tested to be true using only one month of Fraser's data. It implies that the short time interval between the first recording and the overregularization error does not affect the testability of the TP. However, the density of the sample has a more substantial impact on whether the TP can be tested. For Fraser's data, the TP was successfully tested on six files, but not three to five files, as shown in TABLE 6.

Table 4: Summary of Corpus used Small Sample Effects testing

		Age	files	words	input words	verbs	input verbs
Age Related	Fraser_{age}	2;4 - 2;5	11	2861	4074	2104	5497
Density Related	Fraser_3	2;3, 2;4, 2;5	3	1616	3304	1362	577
	Fraser_4	2;0, 2;1,2;3, 2;4	4	1148	1495	1160	640
	Fraser_5	2;0x3, 2;1, 2;2	5	1485	1866	1339	757
	Fraser_6	2;0x3, 2;1, 2;2, 2;4	6	1373	3206	1968	945

Table 5: Number of observed total number of verb types, irregular verbs and exponents in Fraser's samples

		U_p	α_p	U_c	α_c	e_p	β_p	e_c	β_c
Age Related	Fraser_{age}	277	0.66	168	0.73	71	0.53	54	0.60
Density Related	Fraser_3	155	0.80	84	0.85	54	0.64	38	0.70
	Fraser_4	131	0.78	91	0.84	43	0.61	36	0.67
	Fraser_5	145	0.79	95	0.8	53	0.63	33	0.61
	Fraser_6	179	0.76	135	0.85	57	0.6	39	0.71

Table 6: Comparison between the cost in Fraser's samples

		N = Children's production (U_c)			N = Parent's input (U_p)		
		T_R	T_N	$T_R \leq T_N$	T_R	T_N	$T_R \leq T_N$
Age Related	Fraser_{age}	42.58	45.48	True	59.31	79.99	True
Dense Related	Fraser_3	26.36	20.75	**False**	41.38	38.27	**False**
	Fraser_4	26.52	22.55	**False**	33.77	33.82	True
	Fraser_5	25.61	24.69	**False**	40.08	36.56	**False**
	Fraser_6	31.33	31.41	True	45.03	46.24	True

4 Discussion

In language acquisition, rule-based learning is very common, such as past-tense acquisition. But what leads to rule learning in the first place? Yang proposed the Tolerance Principle predict when a rule will be productive. He hypothesizes when the cost to process a word in a list with a rule is minimum a rule will be produced. In order to estimate the cost to process a word, he based his calculations on two assumptions: 1) lexical retrieval is a serial search process; 2) the distribution of children's effective vocabulary is Zipfian. Thus, he uses the TP to quantify a threshold number of exceptions (θ) that a learner can tolerate: $\theta = N/ln(N)$. However, the corpus data doesn't support the threshold predicted by the TP. Yang examined Adam's and Eve's data and neither conform to the TP's threshold.

This paper revisted his second assumption in the TP calculation and revised the testing methodology to make it more appropriate for the corpus data. We then used the revised methods to test eight children's past-tense acquisition data (including Adam's and Eve's). Five of the eight children's data support the TP's prediction. We further investigated the other three children's data and explored how sample size affects the testability of the TP. The results showed that the TP requires a relatively densely sampled corpus to be tested true.

What is a "densely sampled corpus"? Based on our empirical data, N needs to be at least over 120 and the e has to be less than 33% of N. However, we don't have any good explanation. The density of the corpus could affect all four variables in the TP formula, the number of types of items N and exceptions e, and the distribution of N and e, thus changing the exponent α and β. The interaction of the four variables and their relationship to the TP worth further investigation.

Appendix

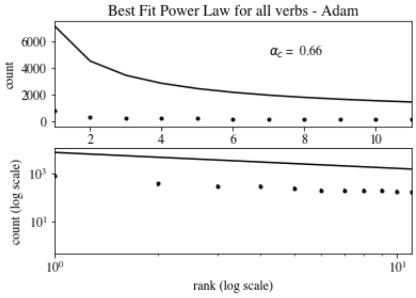

Figure 1: Distribution of Adam's Verbs

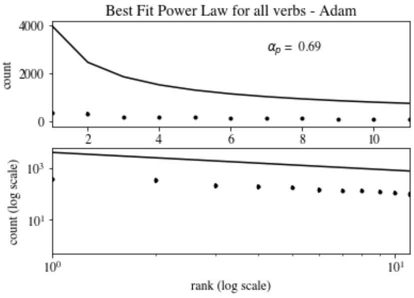

Figure 2: Distribution of Adam's mother's verbs

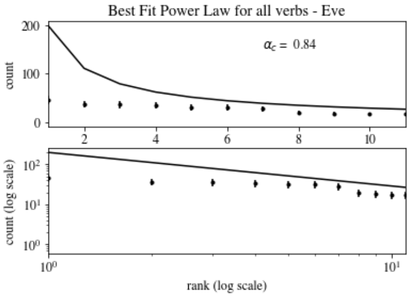

Figure 3: Distribution of Eve's Verbs

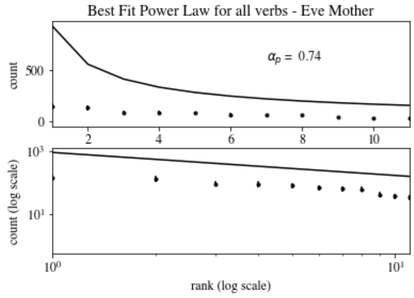

Figure 4: Distribution of Eve's mother's verbs

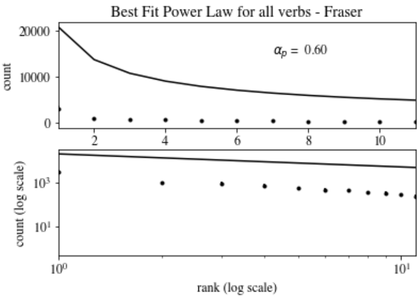

Figure 5: Distribution of Fraser's Verbs

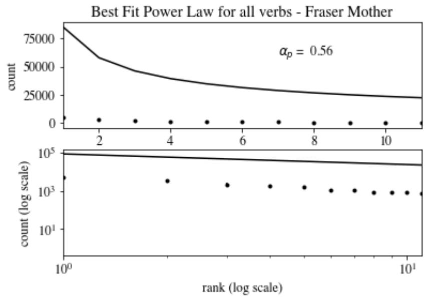

Figure 6: Distribution of Fraser's mother's verbs

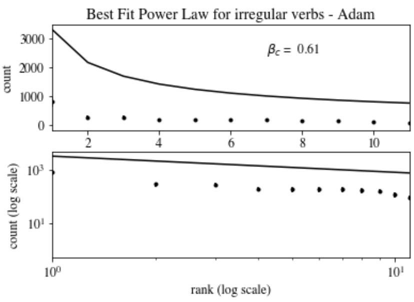

Figure 7: Distribution of Adam's Irregular Verbs

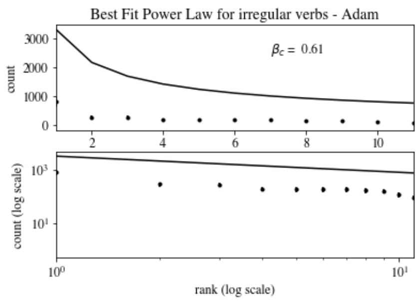

Figure 8: Distribution of Ada's mother's Irregular verbs

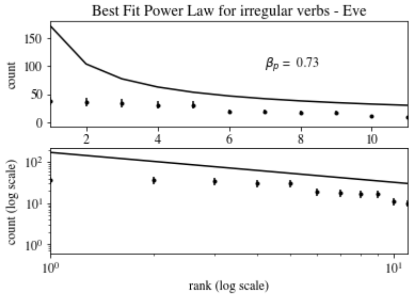

Figure 9: Distribution of Eve's Irregular Verbs

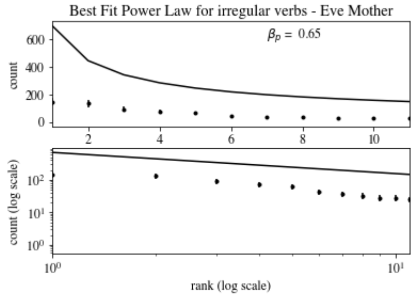

Figure 10: Distribution of Eve's mother's Irregular verbs

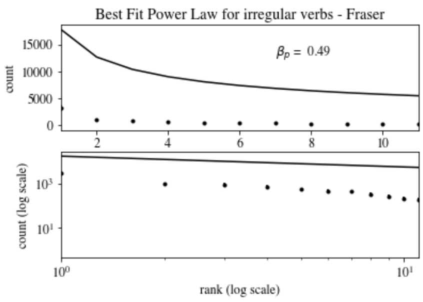

Figure 11: Distribution of Fraser's Irregular Verbs

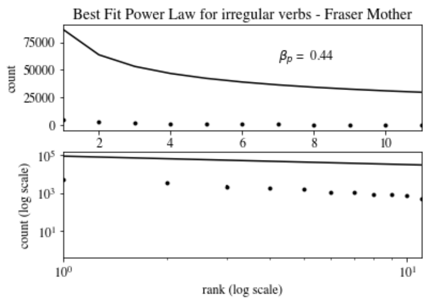

Figure 12: Distribution of Fraser's mother's Irregular verbs

References

John H Andreae and Koenraad Kuiper. Charles yang's tolerance principle. 2018.

Lois Bloom. *One word at a time: The use of single word utterances before syntax*, volume 154. Walter de Gruyter, 1973.

Lois Bloom, Lois Hood, and Patsy Lightbown. Imitation in language development: If, when, and why. *Cognitive psychology*, 6(3):380–420, 1974.

Roger Brown. 1973: A first language: the early stages. cambridge, ma: Harvard university press. 1973.

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

Roy Patrick Higginson. *Fixing: Assimilation in language acquisition*. PhD thesis, Washington State University, 1985.

Vsevolod Kapatsinski. On the intolerance of the tolerance principle. *Linguistic Approaches to Bilingualism*, 8(6): 738–742, 2018.

Elena Lieven, Dorothé Salomo, and Michael Tomasello. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507, 2009.

Brian MacWhinney. Morphosyntactic analysis of the childes and talkbank corpora. In *LREC*, pages 2375–2380, 2012.

Gary F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. Overregularization in language acquisition. *Monographs of the society for research in child development*, pages i–178, 1992.

James L McClelland and David E Rumelhart. An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, 88(5):375, 1981.

Wayne S Murray and Kenneth I Forster. Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, 111(3):721, 2004.

David C Plaut. Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and cognitive processes*, 12(5-6):765–806, 1997.

Jacqueline Sachs. Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's language*, 4:1–28, 1983.

Kathryn D Schuler, Charles Yang, and Elissa L Newport. Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. In *CogSci*, 2016.

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

Charles Yang. Some consequences of the tolerance principle. *Linguistic Approaches to Bilingualism*, 8(6):797–809, 2018a.

Charles Yang. A user's guide to the tolerance principle. *Manuscript. University of Pennsylvania (ling. auf. net/lingbuzz/004146)*, 2018b.