Testing Tolerance Principle on Corpus Data

1 Introduction

1.1 Deriving Tolerance Principle

Yang (2016) in his book proposed Tolerance Principle (TP) to explain how children derive regular forms from noisy input. First he assumed that human brain applies Elsewhere Condition Model (e.g. McClelland and Rumelhart, 1981; Plaut, 1997) in processing rules and exceptions. Elsewhere Condition Model can be implemented as a serial search procedure where the lexical item is evaluated with all the exceptions. If a match is found, exceptional rules is triggered. If not, the rule will be applied. For example, in order to find the correct past tense form for a new item W, W is evaluated with each item (E) in the irregular inventory; if W is not in the irregular inventory, then the general rule will be 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_4 THEN apply rule e_4 to W...
IF W = e_5 THEN apply rule e_5 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...

Based on this model, to retrieve the past tense form for the verb *eat*, *eat* is first evaluated in the irregular inventory and it can be found; therefore the irregular rule for *eat* is applied to derive the past tense form *ate*. To retrieve the past tense form for the verb *type*, it is not found in the irregular set therefore the general past tense rule is applied to derive *typed*.

Given that the lexical retrieval is a serial search process, Yang further proposed that the rules and exceptions are organized to optimize the time complexity in serial search. A productive rule to produce a set of N items is to be derived when rule applying process is less time consuming than processing items individually. When there is no productive rule, all the items are ranked by their frequencies. Therefore, it is always easier to process a more frequently used item. When a productive rule is generated, all items are separated into two categories, regulars and exceptions. The exceptions are ranked by frequencies and the rule is going to apply only when the item is not found in the set of exceptions. For example, for a vocabulary inventory with n regular items (w) and m irregular items (e), when there is no productive rule, 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) No 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 time complexity to process the list with rules is more than without rules. As shown in (2) and (3), there are less items (n+1) to process in the list with productive rule than the list without productive rule (n+m). However, it doesn't necessarily make the time complexity to search for one item in the list with rules less. For example, for the item w_2, which is a more frequent regular word, it's going take less time to find w_2 in the list without productive rule than the list with a rule. Since w_2 is the third most frequent word, one only needs to exhaust all the tokens in w_1 and e_1 in order to find a match for w_2. However, in the list with productive rules, the regular item will be reached after all the tokens in the exceptions are checked, which

creates more time complexity to find w_2. Therefore, only when the average time complexity to search for each word in the list with productive rules (T_R) is less than the list without rules (T_N).

(4) $T_N > T_R$

According to Yang, the average time complexity to search for a word is the product of the probability of a word and the retrieve time for a word. To calculate the probability for each word, Yang assumed that any sample of large corpus follows Zipfian distribution, therefore the product frequency of the word (f_i) and the rank (r_i) 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) over the sum of the frequency of all words. The probability of occurrence (p_i) for w_i can be expressed as (6) where $H_N = \sum_{k=1}^{N} \frac{1}{r_k}$.

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

Since all the items are stored in a list ranked according to their frequency, the retrieval time for each item is strictly based on the rank of the item. The higher the item ranks, the easier it is going to be retrieved. For example, the word *the* appears more frequently in the corpus than the word *give*, therefore it takes less cognitive resource to retrieve the word *the* than *give*. Yang simplified this as 'the *i*-th ranked item takes *i* units of time to be retrieved' (Yang, 2018b, p.???). Therefore, the time complexity for w_i (T_{w_i}) is shown in (7). The total time complexity for N items in the list (T_N) is shown in (8).

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

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

If a rule is generated, all the exceptions are stored in ranked list and all the regular items are stored in a set under the list of exceptions. It is the same process to access all the items in the exception list as all the items in a list without rules. If there are e items of exceptions, the total time complexity (T_{e_i}) for the exception list is shown in (9). The total time complexity for all the items in the exception list (T_e) is shown in (10).

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

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

For the regular items, they can only be accessed after the processing for all the exceptions are done, since a new item is going to be compared to all the exceptions in the list first and if it not in the list the rule will apply. Therefore the time process for all regulars are the same, which is the constant e, given there are e items in the exception list. The total time complexity to process all the regular items ($T_{\overline{w}}$):

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

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

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

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 $T_R \leq T_N$:

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

Since $\frac{e^2}{N} \cdot (\frac{1}{lne} - 1)$ is always smaller 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 (not) it work?

As elaborated in the first part, TP is derived based on Elsewhere Condition Model that the lexical access time is roughly logarithmic (Murray and Forster, 2004). There is no guarantee that it is executed as a cognitive function of learning. TP is appealing since it captures some facts about language acquisition: children generate rules based on noisy input. TP provides an elegant and succinct way to quantify the 'noise' in the input.

Schuler et al. (2016) provided evidence from artificial language learning to support TP. In their study, children between the age of 5 to 7 were presented with nine novel objects with names. The experimenter produced both the singular form and the plural form of the object. They assigned each noun a plural marker that either followed the rule (add *ka*) or not followed the rule with individual suffix (add *po, tay, lee bae, muy,* or *woo*). In one condition, they produced five nouns with *ka* marker and four with individual marker. In another condition, they produced three nouns with *ka* marker and six with individual marker. TP predicts that children will be able to learn under the 5/4 condition but not the 3/6 condition. The results of the experiment confirmed this prediciton, that all children were able to use *ka* as a general plural marker in a Wug-like test.

However, the success of one experiment can not prove that TP to be true. Many researchers have challenged how well it can be applied to explain language acquisition in real life.

Andreae and Kuiper (2018) pointed out that $\log_2 N$ is a better approximate and '$\log_2 N$ is the number of binary features needed to represent N different entities'. Yang used $\ln N$ as the approximate sum for Harmonic number, as shown in (13). For smaller N, $\log_2 N$ is a better approximate. For example in Schuler et al. (2016)'s experiment, when N = 9, $H_9 \approx 2.83$, while $\ln 9 \approx 2.20$ and $\log_2 9 \approx 2.71$. Obviously 2-based logarithm is a better approximate than natural logarithm for N = 9. Moreover, the 4 exceptions was calculated based on $\theta = N/lnN$ where 9/2.20 $\approx$ 4.10. However, if the original $H_9 \approx 2.83$ was inserted, then $\theta = N/H_N$ would be 9/2.83 $\approx$ 3.18. This result will produce a tolerance threshold of three exceptions, which goes against the results in their experiment.

TP also received criticisms on a more general level of how to approach acquisition. Goldberg (2018) criticized that TP only focus on the quantity of the examples but ignored the quality of the examples. She emphasizes the importance of communicative needs, context, prior learning and cognitive load in acquisition. The productive rule construction will be constrained 'when there exists a conventional alternative way to express our intended meaning.' Kapatsinski (2018) criticized the cognitive foundation of TP, serial search. He argued that serial search is a flawed model and a more relevant lexical retrival model should be adopted to derive TP. [1]

In this paper, we agree with Yang's approach in general. However, using corpus data to test if TP can be applied to explain acquisition. Second, to explore a method to use corpus data to test such principle.

2 Testing Tolerance Principle on Corpus Data

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

In Chapter 4 of Yang (2016)'s book, he established a routine for the execution of the TP. In our study, we are also going to follow his routine.

(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 $\theta_N = N/lnN$ to determine productivity.

Yang first applied this routine in explaining the acquisition of past tense in English. English speaking children usually start to produce 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 overregularization error can be seen as an unambiguous mark for the presense of productive 'add *-d*' rule for past tense.

Adam produced the first overregularization error at the age of 2;11, when he said *What dat feeled like?* (Brown, 1973). This error implied that Adam already constructed the past tense rule. According to TP, Adam must have a

[1] See Yang (2018a) for his defense.

large enough corpus of verbs (N) that the irregular verbs (e) could be tolerant, and e is smaller than $\theta = N/lnN$. 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 the past tense of the verbs, he counted all forms of verbs. According to him, as long as Adam produced one form of the 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 children can learn the rule when there are fewer than 53 irregular verbs out of 300 verbs. However, Yang counted 57 irregular verbs in Adam's total 300 verb lexicons. He attributed the difference between 57 and 53 to the sampling effects.

Furthermore, Yang also 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 1;10 and 1;6, when Eve had her first recording. When $N = 163$, $\theta = N/lnN \approx 32$, which means Eve can only tolerant 32 irregular verbs in order to derive the past tense rule. However, Yang found 49 irregulars in her production, which is again higher than the prediction. He then attributed the difference to the undersampling of Eve's data.

2.2 Revised Testing Methodology

In Yang's test, TP failed to account for Adam's and Eve's data on past tense acquisition. With the proposed new methodology, we aim to preserve Yang's insight but develop a different version of TP.

First, we aim to better estimate the effective vocabulary (N) of the child. In Yang's method, he used the raw count of the verbs children produced as the N. There are two problems with this estimation. First, children's production only reflects part of their effective vocabulary that children don't utter all the words they know. Second, the estimated effective vocabulary should not be an exact number; instead, it would be more accurate to use a possible range for N. We propose to estimate the N through parents' input (U_p) and children's production (U_c), both of which can be extracted from the corpora. Since children do not absorb parents' input completely, and since their productions do not represent their entire linguistic knowledge, we introduce λ to represent comprehension cost (%) and δ to represent production loss (%). λ and δ should range between 0% - 100%, which can also be roughly estimated through the corpora by calculating the overlap rate of parents' vocabulary and children's vocabulary. In addition, in order to compensate for the loss of the data due to undersampling, we introduce X_c and X_p for the words that are not caught in recording sessions for the child and the parents respectively. The estimated N can be written as below:

$$N = (U_p + X_p) \cdot \lambda = \frac{(U_c + X_c)}{\delta} \tag{16}$$

Second, we want to better measure the true distribution of N. Yang assumed Zipfian distribution for all items N and all the exceptions e. However, Zipfian distribution is not guaranteed for small corpus such as all the verbs a 2-year-old child knows. Not having a perfect Zipfian distribution will make formula (5) invalid, thus affects all derivations following it. Recall formula (5): when a corpus follows Zipfian distribution, the product of the frequency of the word (f_i) and the rank (r_i) is a constant C, shown in (17) again. This is derived from the formal expression of Zifp's law, which is shown in (18): the frequency of the rth most frequent word is inversely proportional to its rank. Zifpian distribution is a special case of power law function where the exponent (α) equals to 1. Formula (5) is only valid when α is 1, which fits the Zifpian distribution. When smaller corpus (such as children's effective vocabulary of verbs and irregular verbs) doesn't have a Zifpian distribution and the exponent (α) is not 1, formula (5) is not valid anymore, thus all the following derivations will not hold true.

$$r_i \cdot f_i = C \text{ (replicate of (5))} \tag{17}$$

$$f_i = C r_i^{-\alpha}, \alpha = 1 \tag{18}$$

In this paper, we propose to measure the real distribution of all the verbs (N) and irregular verbs (e), and use the empirically estimated exponent in calculation. Since the distribution of all the verbs and irregular verbs are not necessarily the same, we are going to use α and β for the exponent respectively. To include the exponent as a variable, formula (6) to (13) can be rewritten as follows:

(19) Probability of occurence for ith ranked word (p_i):

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

[2]Yang might have made an error in his book. Eve made the first overregularization error at the age of 1;8, when she uttered *I seed it* meaning *I saw it.*

(20) Time complexity for a list of N items without rule (T_N):

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

(21) Time complexity for a list with e exceptions and productive rules (T_R):

$$T_R = \frac{H_{e,\beta-1}}{H_{e,\beta}} \cdot \frac{e}{N} + (1 - \frac{e}{N}) \cdot e$$

(22) 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 foumula (13) where H_N can be conveniently approximated using lnN, there is no mathematical approximation for the Harmonic number in the inequation (22). Therefore, inequation (22) can not be used to estimate the value of θ, the maximum number of irregular verbs children can tolerant before they learn the past tense rule. Instead of comparing the estimated number of irregular verbs and observed irregular verbs in children's vocabulary, we propose to compare T_R and T_N directly. In the next section, we are going to extracted all the variables (e, N, α and β) from three children's corpora to compare T_N and T_R. Tolerance Principle will be confirmed if T_R is smaller than T_N as predicted in (22).

2.3 Testing on Adam's, Eve's and Fraser's Data

In this section, we are going to use the revised testing methods to test Adam's, Eve's (Brown, 1973) and Fraser's data (Lieven et al., 2009) on their past tense acquisition. As mentioned in the previous paragraph, Adam's recording starts at the age of 2;3 and he made the first overregularization error at the age of 2;11. Eve's recording starts at the age of 1;6 and she made the first overregularizaiton error at the age of 1;8. Fraser's recording starts at 2;0 and he made the first overregularization error at the age of 2;5 (Maslen et al., 2004). A summary of their corpus data is shown in table 1.

Table 1: Summary of Adam's, Eve's and Fraser's Data

	Adam	Eve	Fraser
Age of first recording	2;3	1;6	2;0
Age of first overregularization error	2;11	1;8	2;5
Overregularization error	*What dat feeled like*	*I seed it*	*and Grandma seed the fish*
No. of files in between	18	5	90

All of the data were automatically extracted from the annotated corpora in CHILDES using NLTK python package. The verbs in each file were identified using part-of-speech taggers annotated by MOR program (MacWhinney, 2012). The number of verbs and irregular verbs in parent's input (U_p and e_p) and in children's production (U_c and e_c) is shown in table 2. The distribution of the all the verbs and irregular verbs were fitted onto power law distribution. The exponent for all the verbs in parents input (α_p), in children's production (α_c) and the exponent for all the irregular verbs in parents' input (β_p) and children's production (β_c) are shown in table 3. The log graph can be found in appendix.

Table 2: Counts of verbs and irregular verbs

	Adam	Eve	Fraser
U_p	275	136	566
U_c	270	91	358
e_p	70	50	97
e_c	62	36	78

Table 3: Exponent for verbs and irregular verbs

	Adam	Eve	Fraser
α_p	0.69	0.74	0.56
α_c	0.66	0.84	0.60
β_p	0.64	0.65	0.44
β_c	0.61	0.73	0.49

Appendix

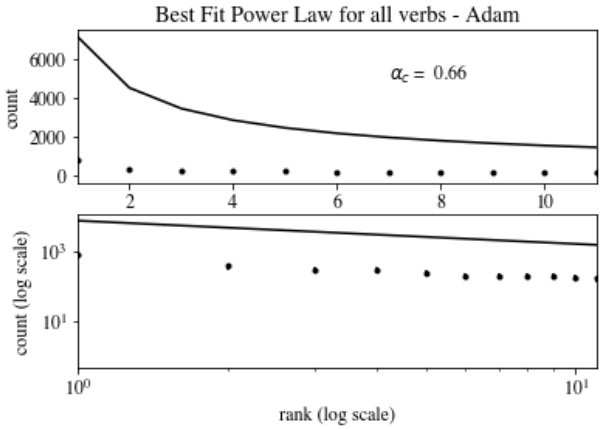

Figure 1: Distribution of Adam's Verbs

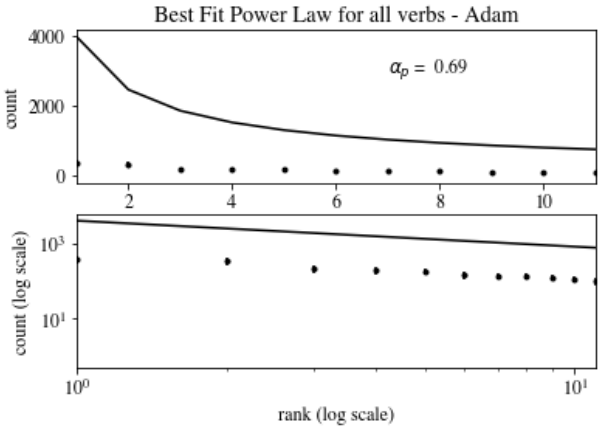

Figure 2: Distribution of Adam's mother's verbs

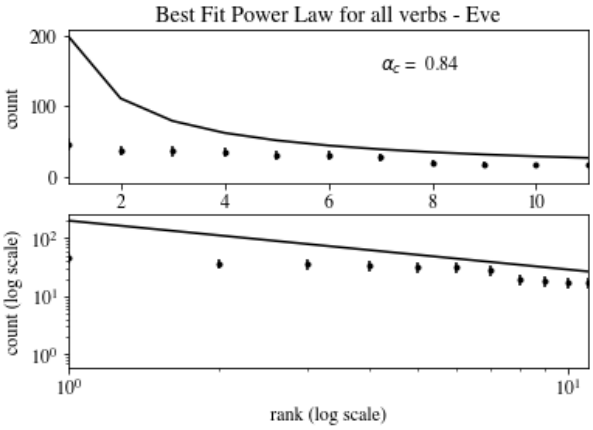

Figure 3: Distribution of Eve's Verbs

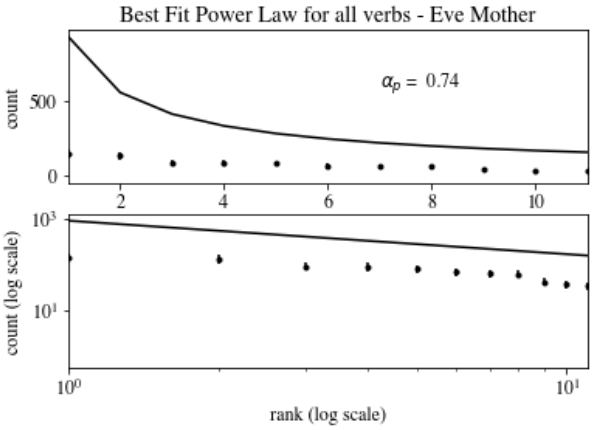

Figure 4: Distribution of Eve's mother's verbs

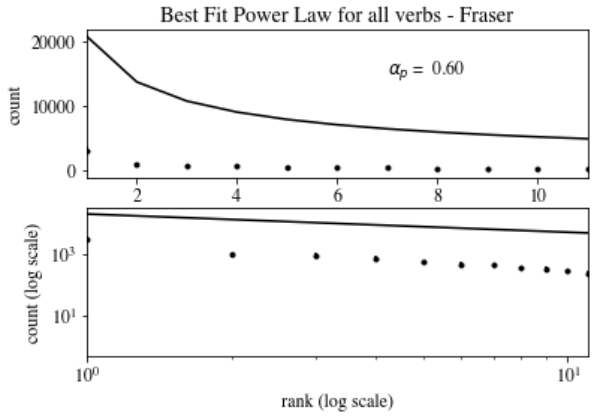

Figure 5: Distribution of Fraser's Verbs

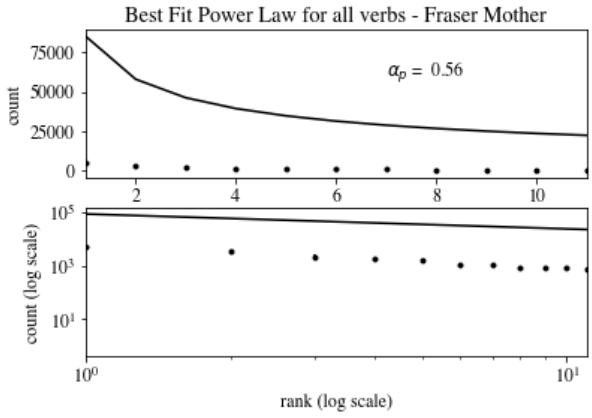

Figure 6: Distribution of Fraser's mother's verbs

References

Andreae, J. H. and Kuiper, K. (2018). Charles yang's tolerance principle.

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

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

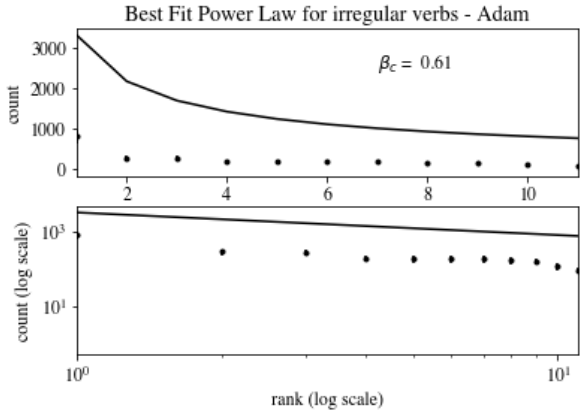

Figure 7: Distribution of Adam's Irregular Verbs

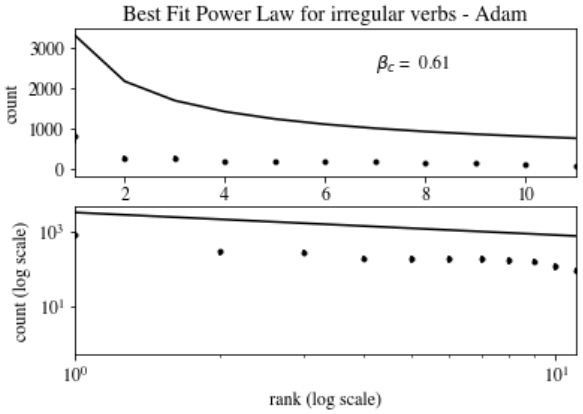

Figure 8: Distribution of Ada's mother's Irregular verbs

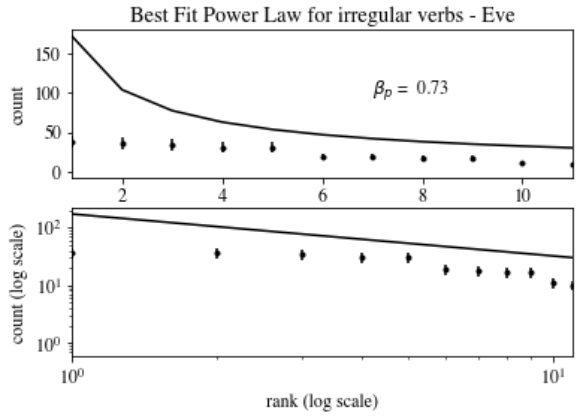

Figure 9: Distribution of Eve's Irregular Verbs

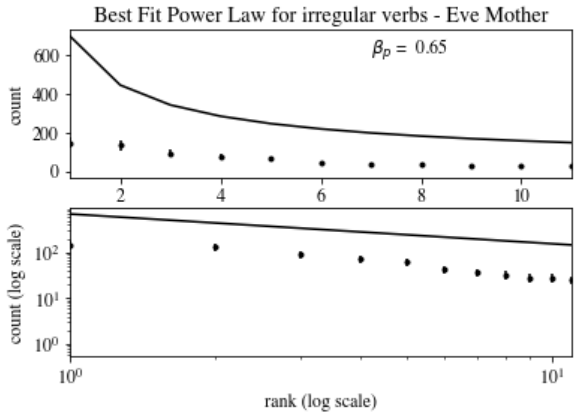

Figure 10: Distribution of Eve's mother's Irregular verbs

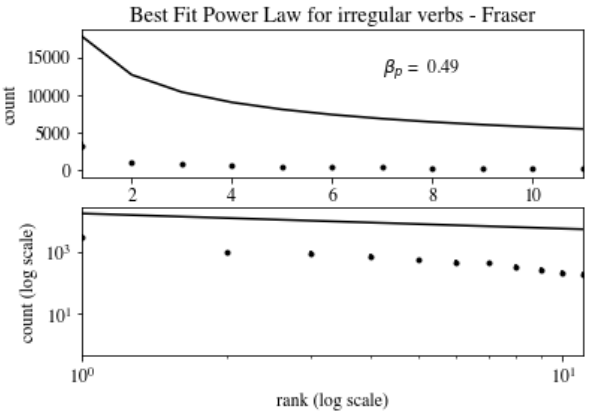

Figure 11: Distribution of Fraser's Irregular Verbs

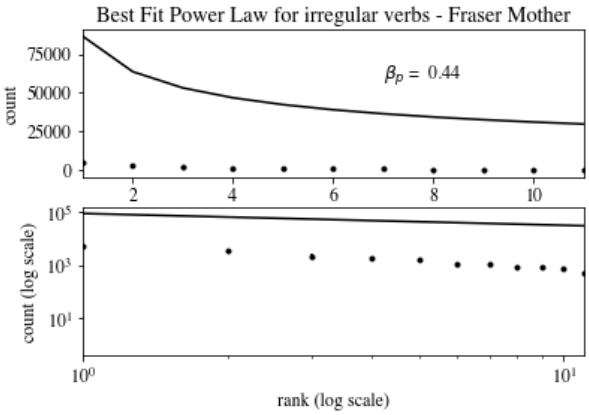

Figure 12: Distribution of Fraser's mother's Irregular verbs

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

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

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

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., and Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development*, pages i–178.

Maslen, R. J., Theakston, A. L., Lieven, E. V., and Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in english. *Journal of Speech, Language, and Hearing Research*.

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

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

Plaut, D. C. (1997). 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.

Schuler, K. D., Yang, C., and Newport, E. L. (2016). Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. In *CogSci*.

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

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

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