Testing Tolerance Principle on Corpus Data

1 Introduction

1.1 Deriving the Tolerance Principle

Rule-based learning, such as past-tense acquisition, is very commonly observed in language acquisition, as when children form the regular past tense for a novel verb in an experimental setting (*wug -> wugged*) (e.g. Berko, 1958) or when they spontaneously produce an overregularization of an irregular verb (*go -> goed*) (e.g. Marcus et al., 1992; Yang, 2000; Pinker and Ullman, 2002). Such evidence indicates that the rule is productive since it outputs word forms that children have not previously encountered in their input. But what leads to use of rules in the first place?

Yang (2005, 2016a) proposed the Tolerance Principle (TP) to predict when a productive rule will be deployed by the language learner. In principle, a productive rule should be derived when it delivers a more efficient result. In this context, Yang used lexical access time to measure efficiency. He hypothesized that a productive rule reduces the total time complexity to access all lexical items.

To estimate the time complexity, Yang first chose serial search as the lexical access model. Serial search model (e.g. Forster, 1976, 1992) proposes that there are two stages in lexical access. The first stage involves a sequential search of all the lexical entries, which is a ranked list by their frequency. When a match is found, the matched lexical entry is used as an index to retrieve all the necessary information 'from a separate file' (Murray and Forster, 2004), such as its semantics, its orthography or its past tense form. For example, in a vocabulary inventory with m rule-following items (w) and n exception items (e), all the items are ranked by their frequency. To access certain lexical information for word w_i, first w_i is compared with each item in the lexical entry starting from w_1. When a match is found, the search terminates automatically and the lexical information is retrieved. The search procedure is shown in TABLE 1.

Applying a productive rule to the general serial search (GSS) model should reduce the time complexity. Yang assumed Elsewhere Condition in rule application. Elsewhere Condition (e.g. Anderson, 1969; Kiparsky, 1973; Halle and Marantz, 1993) proposes that a more specific form is preferred over a more general one when both are available. When Elsewhere Condition is implemented in the serial search procedure, the items that don't follow the productive rule are given priority in the search. Thus, Yang built a rule-based lexical access model combining serial search and Elsewhere Condition, which he called Elsewhere Condition Serial Search (ECSS). In the ECSS model, all the exception items (e) are stored in a ranked list by frequency, and all the rule-following items are concatenated into a set. To access certain information (e.g. past tense form) of a word (w_i), w_i is first compared to the exception list. If a match is found,

a specific form for the matching exception is retrieved. If not, the productive rule is applied. The search process for ECSS is shown in FIGURE (2). For example, w_{e_4} can be irregular verb *eat* and w_j can be regular verb *type*. To access the past tense form for *eat* and *type*, the two verbs are first compared to the irregular verbs in Bin A. When a match is found for *eat*, the past tense form *ate* is retrieved. When the search is exhausted in Bin A and still no match for *type*, the general past tense rule is applied to *type*, outputting *typed*.

Figure 1: General Serial Search Model

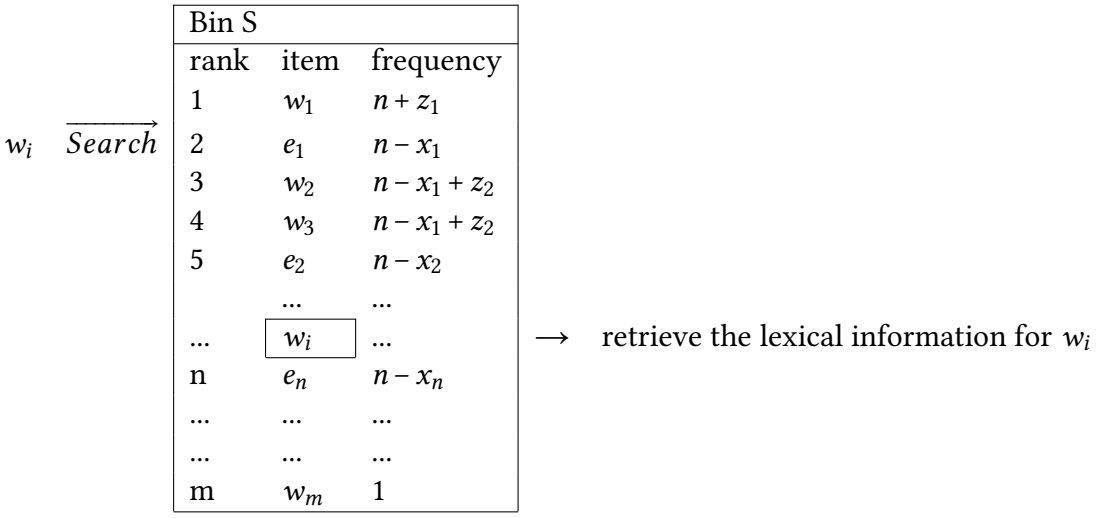
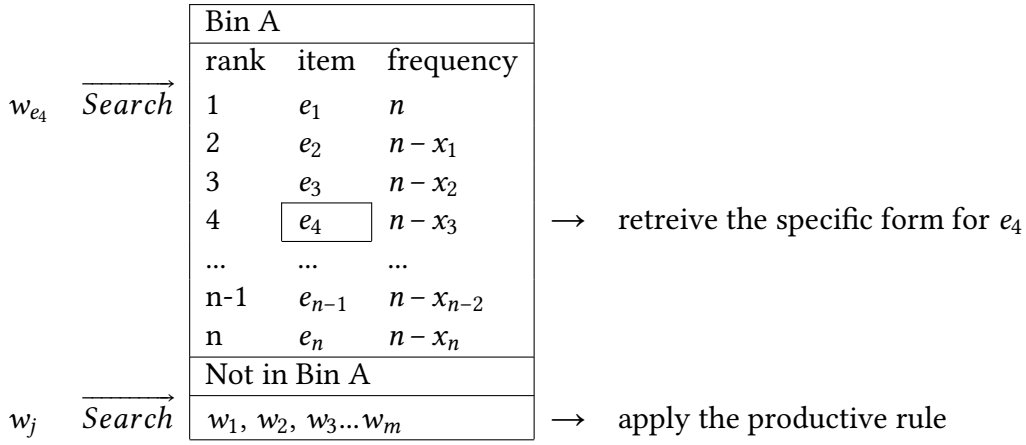


Figure 2: Elsewhere Condition Serial Search Model



Judging on the basis of FIGURE 1 and 2, the ECSS model seems to be a more efficient search procedure since there are fewer items in the lexical list. However, occasionally it could be more time consuming to search for an item in the ECSS model. For example, for the item w_2, it will take less time to find it in the GSS model. In the GSS model, since w_2 is the third most frequent word, the search can be terminated after examining w_1 and e_1. However, in the ECSS model, the rule-following item w_2 will be reached only after all the exceptions are checked, which creates more time complexity. Therefore, the time complexity for the ECSS model is determined by the number of exceptions and the location of the exceptions in the list. Based on Yang's hypothesis, the TP hypothesis will hold true when the time complexity for ECSS model (T_E) is less than that of the GSS model (T_G).

The average time complexity for each word is product of the probability of the word (p_i) and the access time (t_i). For a list of N words, the total time complexity for GSS model (T_G) is:

(1) $$T_G = \sum_{i=1}^{N} (p_i \cdot t_i)$$

When a productive rule is derived, the ECSS model will be in use. The total time complexity for the ECSS model involves two parts, the time complexity for the exception list and the time complexity for rule-following item set. The total time complexity for rule-following items (T_w) is the time complexity, which is the product of an exception item's probability p_j and the lexical access time t_j, and the time complexity for the rule-following item set, which is the product a rule-following item's probability p_k and the lexical access time t_k. The total time complexity (T_E) is shown in (2).

(2) $$T_R = \sum_{j=1}^{e} (p_j \cdot t_j) + \sum_{k=1}^{N-e} (p_k \cdot t_k)$$

Based on Yang's hypothesis, the TP will be true if T_E is smaller than T_G. Therefore, the TP is derived:

(3) Tolerance Principle:
Let T_G and T_E be the lexical access time complexity for a ranked list of items without a productive rule and with one. A productive rule is derived if and only iff:

$$T_G \geq T_R$$

1.2 The TP and the Exceptions

In real life rule-based learning, children encounter exceptions very often, if not more often than the rule-following items. For example, the most frequently used verbs in child directed speech are more likely to be irregular verbs, *go, have, do, eat, drink....* Children can only generate rules when the input has enough rule-following items. Therefore, the rule-based learning question can also be framed as: How many exceptions can a child tolerate in order to derive the rule? Since the number of exception type (e) is a variable in the time complexity formula for the ECSS model (T_E), Yang used the TP inequation to estimate the maximum value for (e):

(4) $$\sum_{i=1}^{N} p_i \cdot t_i \geq \sum_{j=1}^{e} (p_j \cdot t_j) + \sum_{k=1}^{N-e} (p_k \cdot t_k)$$

To calculate the probability for each word (p_i) in T_G, Yang assumes that the ranked word list in GSS model follows the Zipfian distribution; therefore, the product of the word's frequency (f_i) and its rank (r_i) is a constant (C):

(5) $$r_i \cdot f_i = C_i$$

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

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

The lexical access time (t_i) is generally believed to have a logarithmic relation to its frequency (e.g. Howes and Solomon, 1951; McCusker, 1977; **?**). This implies that the access time is not determined by the absolute frequency, but a transformed function of the frequency. Murray and Forster (2004) further proposed the *rank hypothesis* arguing that the access time is directly related to the rank position of a frequency-based ranked list. Yang adopted the *rank hypothesis* and simplified it as 'the i-th ranked item takes i units of time to be retreived' (Yang, 2018), shown in (7).

(7) $t_i = r_i$

Inserting p_i and t_i to the formula in (1), T_G can be written as:

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

As for the T_E, since the exceptions and rule-following items are stored in a ranked list and a set separately in the ECSS model, the probability for the exception item (p_j) and the rule-following item (p_k) are different. The probability of the exception (p_j) the product of the probability of the exception list in the total vocabulary inventory ($Pr(exception)$) and the probability of the exception item in the list ($Pr(j)$). Yang argued that $Pr(exception)$ is the types of exceptions (e) divided by the types of all items (N), as shown in (9a). He assumed that after a rule is derived, the learner can already distinguish a rule following item and an exception, which means the token of each item doesn't play a role. Therefore, $Pr(exception)$ is calculated as the exception types over total types instead of tokens. To estimate $Pr(j)$, Yang assumed that the items in the exception list also follows Zipfian distribution. Therefore, equation (5) can also be used to calculate $Pr(j)$, which is shown in (9b). The probability of an exception item (p_j) is the product of $Pr(exception)$ and $Pr(j)$, which is shown in (9c). The lexical access time for each exception item t_j equals the rank of the item r_j.

(9) a. $$Pr(exception) = \frac{e}{N}$$

b. $$Pr(j) = \frac{f_j}{\sum_{k=1}^{e} f_k} = \frac{1}{r_j \cdot H_e}$$

c. $$p_j = Pr(exception) \cdot Pr(j) = \frac{e}{N} \cdot \frac{1}{r_j \cdot H_e}$$

d. $$t_j = r_j$$

As for rule-following item, the probability and lexical access time for all the rule-following items should be the same. Yang assumed that the lexical access time (t_k) equals to e, since there are e exceptions in the list, as shown in (10a). The sum of p_k is the probability of the rule-following item set in the total vocabulary inventory, which is shown in (10b)

(10) a. $$t_k = e$$

b. $$\sum_{k=1}^{N-e} p_k = 1 - Pr(exception) = 1 - \frac{e}{N}$$

Inserting p_j, t_j, t_k and $\sum_{k=1}^{N-e} p_k$ back to equation (2), T_E can be written as:

$$(11) \quad T_R = \sum_{j=1}^{e}\left(\frac{e}{N}\cdot\frac{1}{r_j\cdot H_e}\cdot r_j\right)+\left(1-\frac{e}{N}\right)\cdot e = \frac{e}{N}\cdot\frac{e}{H_e}+\left(1-\frac{e}{N}\right)\cdot e$$

Therefore, $T_G \geq T_R$ can be written as:

$$(12) \quad \frac{N}{H_N} \geq \frac{e}{N}\cdot\frac{e}{H_e}+\left(1-\frac{e}{N}\right)\cdot e$$

To calculate maximum number for e, Yang first approximated H_N as lnN. H_N is always larger than lnN: $H_N = \sum_{k=1}^{N}\frac{1}{k} = \ln N + \gamma + \epsilon_N$, where γ is the Euler-Mascheroni constant ($\gamma \approx 0.58$) and $\epsilon_N \sim \frac{1}{2N}$. When N is large enough, the difference between H_N and lnN is smaller than 1. Thus, formula (12) can be written as:

$$(13) \quad \begin{aligned} \frac{N}{lnN} &\geq \frac{e}{N}\cdot\frac{e}{lne}+\left(1-\frac{e}{N}\right)\cdot e \\ \frac{N}{lnN} &\geq \frac{e^2}{N}\cdot\left(\frac{1}{lne}-1\right)+e \end{aligned}$$

Since $\frac{e^2}{N}\cdot\left(\frac{1}{lne}-1\right)$ is always smaller than or equal to zero, as long as e is smaller than $\frac{N}{lnN}$, then the inequation will always hold true. Therefore, Yang summarized another version of the TP with exceptions:

(14) Tolerance Principle (with exceptions):
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 $\theta = \frac{N}{lnN}$ (Yang, 2016b, p.64)

1.3 The Testability of the TP

Instead of using the original version of the TP as in (3) in empirical testing, Yang used (14) as the functional TP. In this version, the TP will hold true if the observed types of exceptions (e) is smaller than than the theoretical threshold $\frac{N}{lnN}$, where N is the type of all items in the corpus. The exception version of the TP is much simpler than original version of the TP where the time complexity is calculated. It is less data-demanding and more user-friendly, since there are only two variables (e and N) in the exception version, comparing to eight variables in the time complexity version. However, the exception version doesn't necessarily represent the original TP, since it is simplified based on many assumptions that are always true in practice, such as corpus follows Zipfian distribution and the lexical access time equals rank. The testing results on the exception version of the TP also provided controversial results.

Evidence from artificial language learning (Schuler et al., 2016) supports the exception version of TP. In the experiment, 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. According to the TP, the exception threshold is $\frac{N}{lnN}$ (N=9), which is 4.10. It means that the children can tolerate up to 5 exceptions out of 9 items in order to derive a rule. As the TP predicts, children learned the rule under the 5-exceptions/4-regular condition but not the 6-exceptions/3-regular condition, as shown by their ability to use *ka* as a general plural marker in a Wug-like test.

However, the corpus data doesn't conform to the prediction of the exception version of the TP. Yang applied it to explain the past tense acquisition on Adam's and Eve's data from Brown corpus (Brown, 1973). 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; Yang, 2000; Pinker and Ullman, 2002). 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?*. This error implied that Adam had already constructed the past tense rule. According to TP, the number of irregular verbs that Adam knew (e) must be smaller than $\theta = \frac{N}{lnN}$, 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 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. Therefore, Adam can learn the rule when there are fewer than $\frac{N}{lnN}(N = 300) \approx 53$ irregular verbs. However, Yang counted 57 irregular verbs in Adam's total 300 verb corpus. 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*[1]. Yang found 163 verbs Eve produced between the age 1;6, when Eve had her first recording, and 1;10. When $N = 163$, $\frac{N}{lnN} \approx 32$, 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.

It's unfair to attribute the TP's failure to the sampling effects given only two children's data. Similarly, it is unfair to conclude the TP doesn't hold true on corpus data. In this paper, we argue that the exception version of the TP is not testable on corpus data. As mentioned above, it assumes that the corpus follows a Zipfian distribution. In an artificial language learning experiment, the corpus can be designed to be Zifpian. However, a Zipfian distribution is not guaranteed for real-life corpus data, such as all the verbs and irregular verbs a 2-year-old child knows. The exception version of the TP also assumes that the word's lexical access time equals its rank location, which might not be true in real-life acquisition. One important prerequisite for the rank hypothesis is that the speaker has to establish a stable frequency-based ranked vocabulary list that each word have a fixed rank position regardless of the increasing absolute frequency. The rank hypothesis also requires the speaker to have perceived familiarity for each word, which is also deviant from the word's absolute frequency. In addition, age of acquisition is also a very important factor in determining the rank position of each word (e.g. Morrison and Ellis, 1995; Ellis and Morrison, 1998; Gerhand and Barry, 1999). In an

[1]Yang made an error here. Eve made the first overregularization error at the age of 1;8 (Brown/Eve/010800.cha), when she said *I seed it.*

artificial language experiment, the perceived familiarity and the age of acquisition of a word are easily controlled. However, in corpus data, there is no way to measure how these two factors impacted the rank position of a certain word. Moreover, the most supportive evidence for the rank hypothesis comes from lexical decision task data in college students, who satisfy all the prerequisites for the rank hypothesis (Murray and Forster, 2004). The corpus data are collected on language learning children. Therefore, the rank hypothesis probably is the most appropriate way to estimate time complexity in corpus data.

In this paper, we propose that the discrepancy of the TP's prediction and the corpus data can't be simply attributed to the sampling effects, and it doesn't show that the TP has failed either. We argue that the exception version of the TP is not testable on the corpus data because many assumptions don't hold true on corpus. To properly test the TP on corpus data, we propose to directly compare the time complexity instead of comparing number of exceptions as a proxy. The rest of the paper is organized as follows: in section 2, we introduce a corpus-testable TP based on the original version in (3). In section 3, we test it on eight children's corpus data, including Adam and Eve. In section 4, we discuss the implications of the results on the TP on corpus data.

2 Testing Corpus

2.1 Corpus Data

We use eight children's corpus data from CHILDES (MacWhinney (2000) to test Tolerance Principle. These children's past tense acquisition have been intensively studied in previous literature[2]. We adopted Yang's data selection criteria, that we included all the files from the first recording file to the file that the child made his/her first overregularization error. The sample size and sample density varies for each child. The average age range is about 8 months from the age of first recording to age of the first overregularization error, with a minimum of 2 months interval (Eve 1;6 - 1;8) and maximum 18 months interval (Allison 1;5 - 2;11). On average, each child has 23.5 files, with a maximum of 90 files (Fraser) and a minimum of 2 files (April).

Table 1: Summary of corpus data for each child

	Age range	corpus	files	child's word tokens	parent's word tokens
Adam	2;3 - 2;11(*feeled*)	Brown (1973)	18	39,403	30,366
Eve	1;6 - 1;8 (*seed*)	Brown (1973)	5	5,304	11,253
Sarah	2;3 - 2;10 (*heared*)	Brown (1973)	33	18,778	27,682
Peter	1;3 - 2;6 (*broked*)	Bloom et al. (1974)	14	52,769	95,180
Naomi	1;3 - 1;11 (*doed*)	Sachs (1983)	20	8,009	9,634
Allison	1;5 - 2;11 (*throwed*)	Bloom (1973)	6	4,605	9,366
April	1;10 - 2;1 (*boughted*)	Higginson (1985)	2	1,376	4,435
Fraser	2;0 - 2;5 (*seed*)	Lieven et al. (2009)	90	137,407	222,200

All of the data were automatically extracted from the annotated corpora in CHILDES using

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

the NLTK python package. The verbs in each file were identified using part-of-speech taggers annotated by the MOR program (MacWhinney, 2012). Instead of only using the verbs in children's production, we also counted all the verbs the parents' produced. Since the N and e represents all verbs and irregular verbs the child knows, the child's production and the parents' input provide the lower and upper bounds of the child's vocabulary. A summary of the verbs and irregular verbs is shown in TABLE 2.

Table 2: Summary of verbs and irregular verbs

	Child's production				Parents' input			
	verb tokens	verb types	irregular tokens	irregular types	verb tokens	verb types	irregular tokens	irregular types
Adam	6,747	306	3,632	62	4,670	297	2,863	70
Eve	564	93	337	36	1,618	138	966	50
Sarah	1,759	189	1,035	48	3,867	293	2,525	68
Peter	7,532	424	3,647	67	15,537	633	8,466	83
Naomi	1,240	128	757	43	1,463	174	945	59
Allison	612	88	335	36	1,453	140	936	44
April	128	50	62	19	658	100	429	37
Fraser	13,924	371	9,903	78	32,359	581	23,169	97

2.2 Corpus-testable TP

The original version of the TP in (3) compares the lexical access time complexity for the GSS model (T_G) and ECSS model (T_E). In this section, we use corpus data to calculate all the variables shown in in the formula in (1) and (2) (repeated in 15 and 16) to calculate T_G and T_E directly. The TP will be tested to be true if $T_G \geq T_E$.

$$T_G = \sum_{i=1}^{N} (p_i \cdot t_i) \tag{15}$$

$$T_E = \sum_{j=1}^{e} (p_j \cdot t_j) + \sum_{k=1}^{N-e} (p_k \cdot t_k) \tag{16}$$

Table 3: Summary of the variables

p_i	probability of a verb in a ranked list	t_i	access time of a verb in a ranked list
p_j	probability of an irregular verb in a ranked list	t_j	access time of an irregular verb in a ranked list
p_k	probability of a regular verb in an unordered set	t_k	access time of a regular in an unordered set

First we would like to reduce some of the variables. Since all the regular verbs are stored in a set, the probability (p_k) and the access time (t_k) should be same for all regular verbs. All the regular verbs are processed after they are compared to all the irregular verbs in the ranked list, therefore t_k should be the total time complexity of all the irregular verbs, which is

$\sum_{j=1}^{e}(p_j \cdot t_j)$. The sum of the p_k is the probability of a word that is not a irregular verb combining the ranked list and the unordered set, which is $1 - \frac{S_e}{S_N}$, S_e and S_N are the tokens for irregular verbs and all verbs. Therefore, the formula for T_E can be written as:

$$T_E = \sum_{j=1}^{e}(p_j \cdot t_j) + (1 - \frac{S_e}{S_N}) \cdot \sum_{j=1}^{e}(p_j \cdot t_j) = (2 - \frac{S_e}{S_N}) \cdot \sum_{j=1}^{e}(p_j \cdot t_j) \quad (17)$$

Therefore, the formula was reduced to two set of variables: the probability for a verb in a ranked list (p_i and p_j) and the access time for a verb in a raked list (t_i and t_j). The following session, we are discuss how to estimate probability and time complexity.

2.3 Estimating Probability

The exact probability for each word can be calculated based on the word tokens in the corpus. However, the corpus was only a small part of ?????, the exact probability in the corpus doesn't truly reflect the probability of each word in the child's vocabulary. To compensate for the limitation of corpus data, the probability can be estimated based on word frequency distribution.

Word frequency distribution follows a puzzling but systematic pattern that there are few very high-frequency words that account for most of the tokens in the text and many low-frequency words. Zipf's law is the most famous empirical description of this pattern, stating that in a given corpus, the frequency of any word $f(r)$ is inversely proportional to its rank r (Zipf, 1935, 1949). Yang assumed the corpus follows classical Zipfian distribution where the exponent $k = 1$, as shown in TABLE 4. However, word distribution in real life corpus is complex that the classic version of Zipfian distribution usually can't best fit the data. One way to improve the fitness is to use a general version of the Zipf's law where the exponent (k) is greater than 1 (e.g. Adamic and Huberman, 2002; Moreno-Sánchez et al., 2016). Moreover, the Zipf's law can also be generalized by 'padding up' the rank by an amount b, which is the Zipf-Mandelbrot law (Mandelbrot, 1965). Empirical data showed that frequency distribution of word within the same categories fits nicely by Zipf-Mandelbrot law (Piantadosi, 2014). In addition, the classic Zipf's law can also be generalized into a log-normal distribution, which is a better fitting model for word frequency distribution (Carrol, 1967; Carroll, 1969; Baayen, 2002). The formula and PDF for each distribution is shown in TABLE 4.

We compared these four distributions on children's and parents' verb and irregular verb distribution. The curve-fitting graph for each child can be found in Appendix. Though each distribution fits the data differently, it's difficult to select a better fit distribution. Judging from the graphs, the General Zipfian distribution and the Lognormal distirbution seem to fit more data than the other two distributions. The goodness of fit was compared between these two distributions using python *powerlaw* package (Alstott and Bullmore, 2014). R is the loglikelihood ratio between General Zipfian and Lognormal distribution, where the positive number indicating that it is more likely to be Lognormal distribution. All the data had a positive R-value, however only 10 out of 32 distributions are significant at $\alpha = .05$ level. A summary of the comparison can be found in TABLE 5. In conclusion, the curve-fitting graph and loglikelihood comparison couldn't provide enough evidence to decide a best-fitting distribution. Since all the distributions predicted the frequency values are not too deviant from the empirical data, in this study we are going to use the exact probability for each word that is calculated through tokens as p_i and p_j.

Table 4: Possible distribution for word frequencies

	$f(x)$	$p(x)$
Classic Zipfian	$f(r) = C \cdot r^{-k}$ $(k = 1)$	$p(r) = \frac{1}{r \cdot H_N}$
General Zipfian	$f(r) = C \cdot r^{-k}$ $(k \geq 1)$	$p(r) = \frac{1}{r^{\alpha} \cdot H_{N,\alpha}}$ $(\alpha \geq 1)$
Zipf-Mandelbrot	$f(r) = C \cdot (r + b)^{-k}$ $(b \geq 0, k \geq 1)$	$p(x) = \frac{1}{(r + b)^{\alpha} \cdot H_{N,b,\alpha}}$ $(b \geq 0, \alpha \geq 1)$
Lognormal	$f(r) = e^{d - m(lnr)^k}$ $(d, m \geq 0, k \geq 1)$	$p(x) = \frac{e^{-((\ln x)^2/2\sigma^2)}}{x\sigma\sqrt{2\pi}}$ $\sigma > 0$

Table 5: Comparison between General Zipf and Lognormal

	child's production		parent's input	
R	verb	irregular verb	verb	irregular verb
Adam	2.52*	1.39	4.47**	0.88
Fraser	0.31	1.10	0.43	1.08
Peter	4.86**	2.02*	4.64**	1.75
Naomi	3.20**	0.04	1.11	1.28
Allison	0.74	0.76	1.16	0.85
April	0.69	0.21	3.31**	0.55
Eve	3.11**	2.22*	3.63**	0.21
Sarah	3.20**	0.19	3.92**	0.40

2.4 Estimating Access time

The frequency of the word is the most potent factor in determining the lexical access time [3](Whaley, 1978). rd access The lexical access time is related to the frequency of the word. In addition, the relationship of frequency and the access time is logarithmic. The nature of the frequency effects in lexical access is still under debate. Under the lexical retrieval model, frequency is treated as an diagnostic indicator for sequential search (Becker, 1979; Forster, 1976; Paap et al., 1982). Murray and Forster (2004)'s rank hypothesis was also proposed with the assumption that the frequency is an indicator for lexical access time. They further proposed that the optimal procedure would be that the frequency offers the rank position, which is the lexical access time.

However, most of the lexical access time studies are based on adults data, who already have the optimal procedure for lexical search. Very little is known about children's lexical retrieval behavior. In this paper, we are going to assume that frequency plays a bigger role in children's lexical retrieval than simply provides a rank position. Moreover, the frequency

[3]Other factors such as length, regularity, homophony, number of meanings only have an influence when frequency is controlled

does not have a strict inverse porportion to the lexical access time, since the children already acquired the verbs. Therefore, the children's lexical access time should be larger than the optimal rank hypothesis, and smaller than the absolute frequency effects. Since too little is known to estimate the access time for children, we are going to use the rank and absolute frequency as the minimum and maximum access time for each child. The real access time should fall in the range. To keep the same scale for two methods, we are going to make the most frequent word have 1 unit of lexical access time. In the rank hypothesis, the second most frequent word would have 2 units of lexical access time, which is its rank. In the absolute frequency account, the second most frequeng word would have t units of lexical access time, where t is the proportion of the second word's in the first word, which is $\frac{S1}{S2}$. A summary of the access time (t) predicted by tokens (S) is shown in TABLE 6.

Table 6: Summary of lexical access time

	t
Rank hypothesis (mimimum)	$t_i = r_i$
Real lexical access time	$r_i \geq t_i \leq \frac{S_1}{S_i}$
Absolute frequency (maximum)	$t_i = \frac{S_1}{S_i}$

3 Testing the TP on Corpus Data

3.1 Testable TP

For each child, we are going to test the TP on their own production and on their parents' input. The children's production should provide the lower boundary for the time complexity and the parents' input provides the upper bound. To test the corpus data, we are going to use the tokens to calculate the exact probability for each word, since all the distributions don't make a big difference in the actual data fitting. We are going to use the rank position as the minimum access time and absolute frequency as the maximum access time. The formula for T_G and T_E are shown in (18a) and (18b). If the $T_{G-max} > T_{E-max}$ and/or $T_{G-min} > T_{E-min}$, the TP will be tested to be true.

In this paper, each child has four sets of T_G and T_E comparisons: $T_{G-max,child}$ vs $T_{E-max,child}$, $T_{G-max,parents}$ vs $T_{E-max,parents}$, $T_{G-min,child}$ vs $T_{E-min,child}$ and $T_{G-min,parents}$ vs $T_{E-min,parents}$. We predict that the TP is more likely to be tested true on parents' input data and the minimum version of the time complexity, which means that $T_{G-min,parents} > T_{E-min,parents}$ is more probable.

(18) a. $$T_{G-max} = \sum_{i=1}^{N}(\frac{S_i}{S_N} \cdot \frac{S_1}{S_i}) = N \cdot \frac{S_1}{S_N}$$

b. $$T_{G-min} = \sum_{i=1}^{N}(\frac{S_i}{S_N}) \cdot r_i$$

(19) a. $T_{E-max} = (2 - \frac{S_e}{S_N}) \cdot \sum_{j=1}^{e} (\frac{S_j}{S_e} \cdot \frac{S_1}{S_j}) = (2 - \frac{S_e}{S_N}) \cdot \frac{e \cdot S_1}{S_e}$

b. $T_{E-min} = (2 - \frac{S_e}{S_N}) \cdot \sum_{j=1}^{e} (\frac{S_j}{S_e} \cdot r_i)$

3.2 Testing Results

The testing results are shown in TABLE 7. Six children's corpus data fully conformed to the TP's prediction, that all four sets of comparison are true. The maximum version of the T_G and T_E on Allison's and April's data failed the TP's prediction. The values for T_{G-MAX} and T_{E-MAX} are really close for Allison (9.35 vs 10.15) and April (4.3 vs 5.11) that we believe that such small difference could be attributed to sampling effect. April has the smallest sample of all children that she only had 2 files with total 612 verb tokens and 88 verb types. Allison has the least densed sample that she had 6 files over 18 months of interval. Since all of the children have at least three sets of comparisons tested to be true, we would like to conclude that the results demonstrate that corpus data support the TP.

Table 7: Summary of the Testing Results

	Children's production					
	T_{G-MAX}	T_{E-MAX}	$T_{G-MAX} > T_{E-MAX}$	T_{G-MIN}	T_{E-MIN}	$T_{G-MIN} > T_{E-MIN}$
Adam	36.87	20.29	TRUE	33.80	13.65	TRUE
Eve	7.42	6.74	TRUE	17.51	11.02	TRUE
Sarah	18.05	11.00	TRUE	25.65	11.13	TRUE
Peter	40.98	20.27	TRUE	43.82	12.72	TRUE
Naomi	13.32	10.18	TRUE	19.63	11.10	TRUE
Allison	9.35	10.15	FALSE	18.24	12.25	TRUE
April	4.30	5.11	FALSE	14.64	10.00	TRUE
Fraser	85.24	32.47	TRUE	26.34	8.77	TRUE
	Parents' input					
	T_{G-MAX}	T_{E-MAX}	$T_{G-MAX} > T_{E-MAX}$	T_{G-MIN}	T_{E-MIN}	$T_{G-MIN} > T_{E-MIN}$
Adam	23.59	12.58	TRUE	35.27	15.85	TRUE
Eve	12.45	10.60	TRUE	21.17	12.34	TRUE
Sarah	26.97	12.91	TRUE	33.37	13.73	TRUE
Peter	42.68	14.99	TRUE	47.19	15.64	TRUE
Naomi	12.49	8.88	TRUE	27.12	14.04	TRUE
Allison	10.70	7.07	TRUE	21.32	11.11	TRUE
May	7.60	5.81	TRUE	19.09	12.30	TRUE
Fraser	88.36	26.45	TRUE	30.80	10.37	TRUE

4 Discussion

In this paper, we first developed a version of the TP that is testable on the corpus data. Different from Yang's functional TP that compares the number of irregular verbs, this version of the TP compares the time complexity directly. The TP has been tested to be true on eight children's data. Therefore, we conclude that the corpus data supports the TP's claim, that a

productive rule is derived when it reduces the time complexity of the lexical access time in serial search. However, there are still some fundamental assumptions about the TP that need to be address.

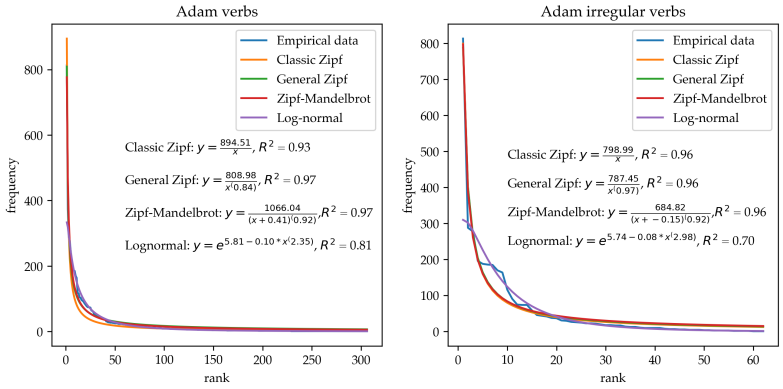
4.1 Serial Search vs Parallel Process

One fundamental assumption of the TP is that the lexical retrieval follows serial search process. However, an alternative to this model could be parallel search, where words are stored in different bins and all the bins are searched in parallel (Forster, 1992). In TP's GSS model, all the items are stored in one ranked list, Bin S. The rank of the item determines the lexical access time. However, if the words are not stored in one ranked list, but in multiple ranked list, and the serial search is conducted on all the ranked lists simultaneously, the lexical access for one word would change drastically. For example, as shown in FIGURE 3, if there is only one list and w_i is the 30th word one the list, the lexical access time for w_i would be 30. However, if there are multiple lists and w_i happens to the the first word in one of the lists, the lexical access time for w_i would be 1.

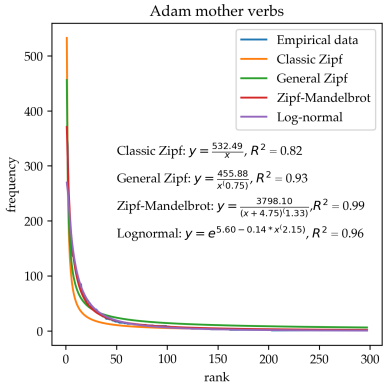
Figure 3: Parallel search for multiple bins

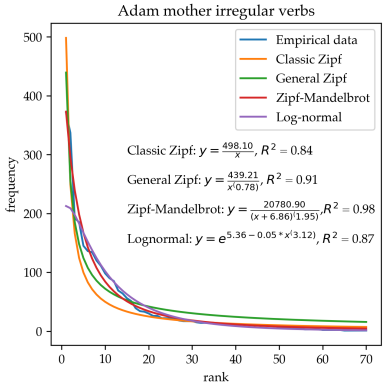
4.2 Static vs Dynamic Learning Process

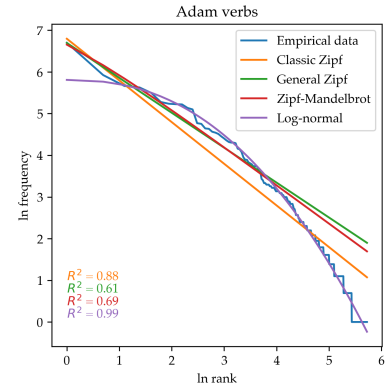
In the calculation of the TP, the rule deriving process was seen as a one-time behavior that when the ratio of irregular verbs and regular verbs reached to an equilibrium the rule will be derived. However, in real life acquisition, the rule deriving behavior should be dynamic, that the comparison of the time complexity of is constantly happening. The acquisition order of the regulars and irregulars are important in the dynamic model. For verb acquisition, most of the commonly used verbs in child directed speech are irregular verbs, which would lead to a late acquisition of the rule. In plural form acquisition, the regular plural forms are also frequent in child directed speech, which would lead to an early acquisition of the plural form. The future work on the TP should be focus on how to develop a dynamic model that could also incorporate the parallel process.

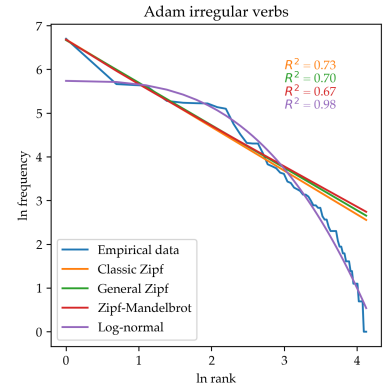

Adam verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{894.51}{x}$, $R^2 = 0.93$
General Zipf: $y = \frac{808.98}{x^{(0.84)}}$, $R^2 = 0.97$
Zipf-Mandelbrot: $y = \frac{1066.04}{(x+0.41)^{(0.92)}}$, $R^2 = 0.97$
Lognormal: $y = e^{5.81 - 0.10 * x^{(2.35)}}$, $R^2 = 0.81$
frequency
rank

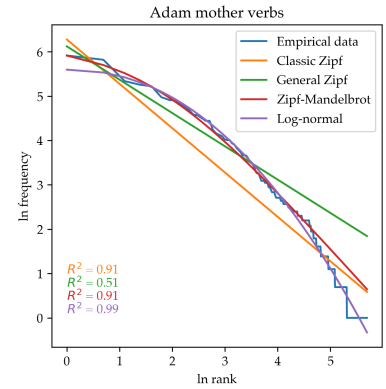
Adam irregular verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{798.99}{x}$, $R^2 = 0.96$
General Zipf: $y = \frac{787.45}{x^{(0.97)}}$, $R^2 = 0.96$
Zipf-Mandelbrot: $y = \frac{684.82}{(x+-0.15)^{(0.92)}}$, $R^2 = 0.96$
Lognormal: $y = e^{5.74 - 0.08 * x^{(2.98)}}$, $R^2 = 0.70$
frequency
rank

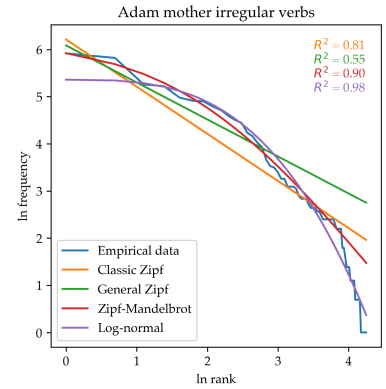

Adam mother verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{532.49}{x}$, $R^2 = 0.82$
General Zipf: $y = \frac{455.88}{x^{(0.75)}}$, $R^2 = 0.93$
Zipf-Mandelbrot: $y = \frac{3798.10}{(x+4.75)^{(1.33)}}$, $R^2 = 0.99$
Lognormal: $y = e^{5.60 - 0.14 * x^{(2.15)}}$, $R^2 = 0.96$
frequency
rank

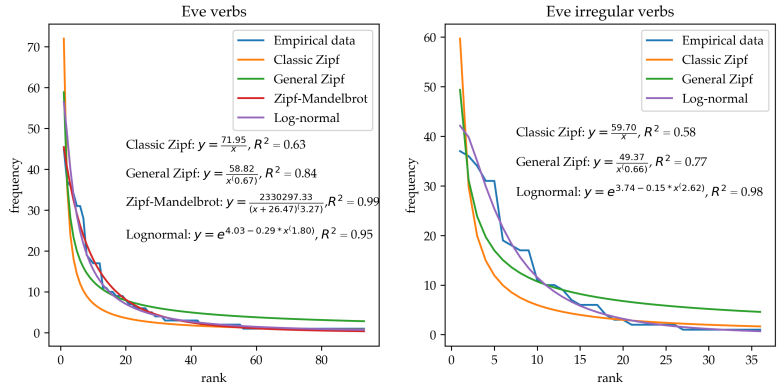

Adam mother irregular verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{498.10}{x}$, $R^2 = 0.84$
General Zipf: $y = \frac{439.21}{x^{(0.78)}}$, $R^2 = 0.91$
Zipf-Mandelbrot: $y = \frac{20780.90}{(x+6.86)^{(1.95)}}$, $R^2 = 0.98$
Lognormal: $y = e^{5.36 - 0.05 * x^{(3.12)}}$, $R^2 = 0.87$
frequency
rank


Adam verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
$R^2 = 0.88$
$R^2 = 0.61$
$R^2 = 0.69$
$R^2 = 0.99$
ln frequency
ln rank

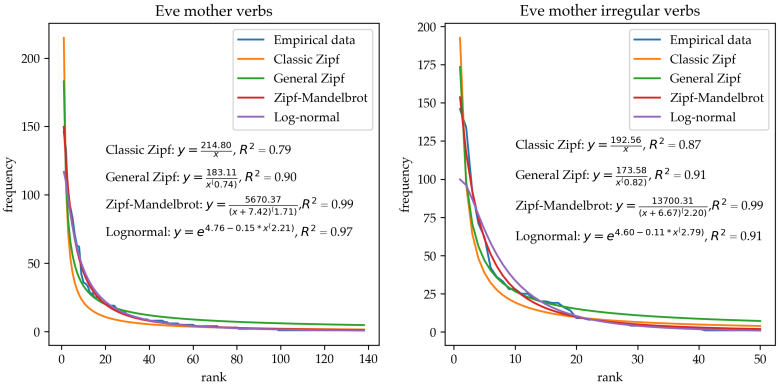

Adam irregular verbs
$R^2 = 0.73$
$R^2 = 0.70$
$R^2 = 0.67$
$R^2 = 0.98$
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
ln frequency
ln rank


Adam mother verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
$R^2 = 0.91$
$R^2 = 0.51$
$R^2 = 0.91$
$R^2 = 0.99$
ln frequency
ln rank

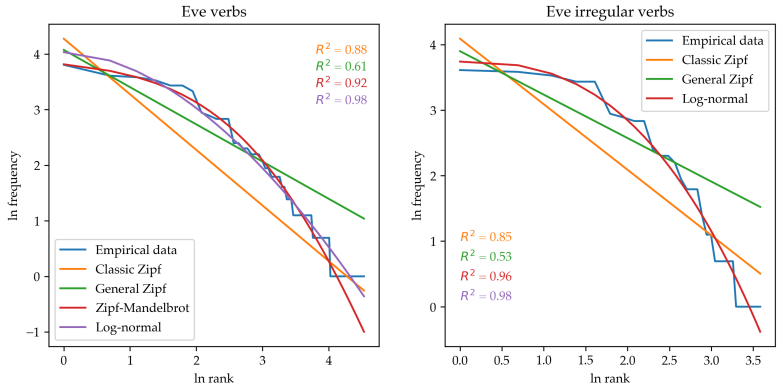

Adam mother irregular verbs
$R^2 = 0.81$
$R^2 = 0.55$
$R^2 = 0.90$
$R^2 = 0.98$
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
ln frequency
ln rank


Eve verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{71.95}{x}$, $R^2 = 0.63$
General Zipf: $y = \frac{58.82}{x^{(0.67)}}$, $R^2 = 0.84$
Zipf-Mandelbrot: $y = \frac{2330297.33}{(x+26.47)^{(3.27)}}$, $R^2 = 0.99$
Lognormal: $y = e^{4.03 - 0.29 * x^{(1.80)}}$, $R^2 = 0.95$
frequency
rank

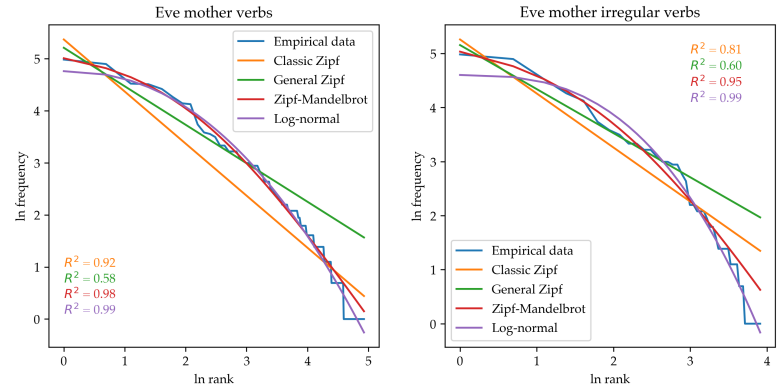
Eve irregular verbs
Empirical data
Classic Zipf
General Zipf
Log-normal
Classic Zipf: $y = \frac{59.70}{x}$, $R^2 = 0.58$
General Zipf: $y = \frac{49.37}{x^{(0.66)}}$, $R^2 = 0.77$
Lognormal: $y = e^{3.74 - 0.15 * x^{(2.62)}}$, $R^2 = 0.98$
frequency
rank


Eve mother verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{214.80}{x}$, $R^2 = 0.79$
General Zipf: $y = \frac{183.11}{x^{(0.74)}}$, $R^2 = 0.90$
Zipf-Mandelbrot: $y = \frac{5670.37}{(x+7.42)^{(1.71)}}$, $R^2 = 0.99$
Lognormal: $y = e^{4.76 - 0.15 * x^{(2.21)}}$, $R^2 = 0.97$
frequency
rank

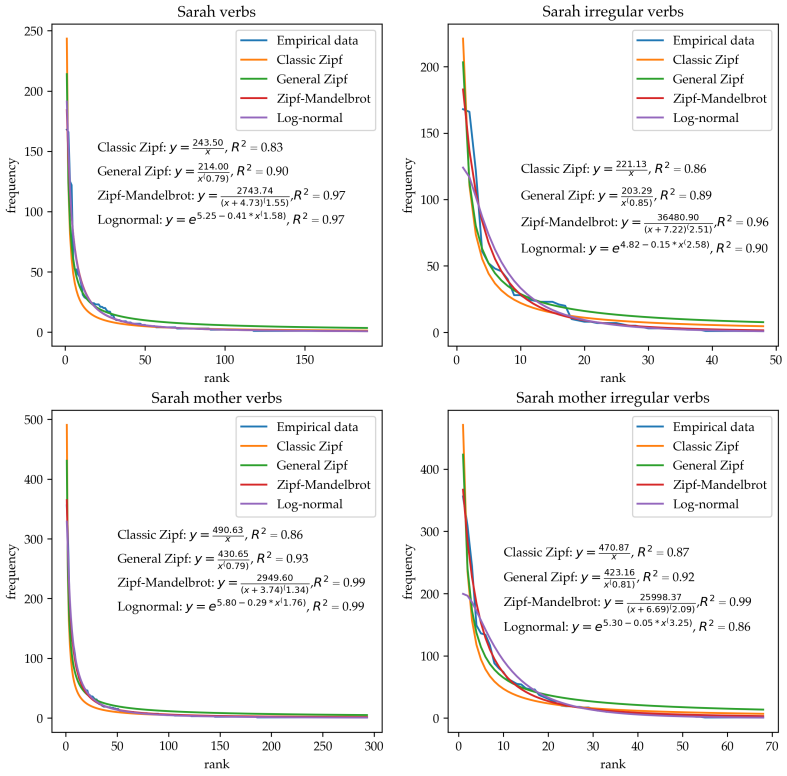
Eve mother irregular verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{192.56}{x}$, $R^2 = 0.87$
General Zipf: $y = \frac{173.58}{x^{(0.82)}}$, $R^2 = 0.91$
Zipf-Mandelbrot: $y = \frac{13700.31}{(x+6.67)^{(2.20)}}$, $R^2 = 0.99$
Lognormal: $y = e^{4.60 - 0.11 * x^{(2.79)}}$, $R^2 = 0.91$
frequency
rank

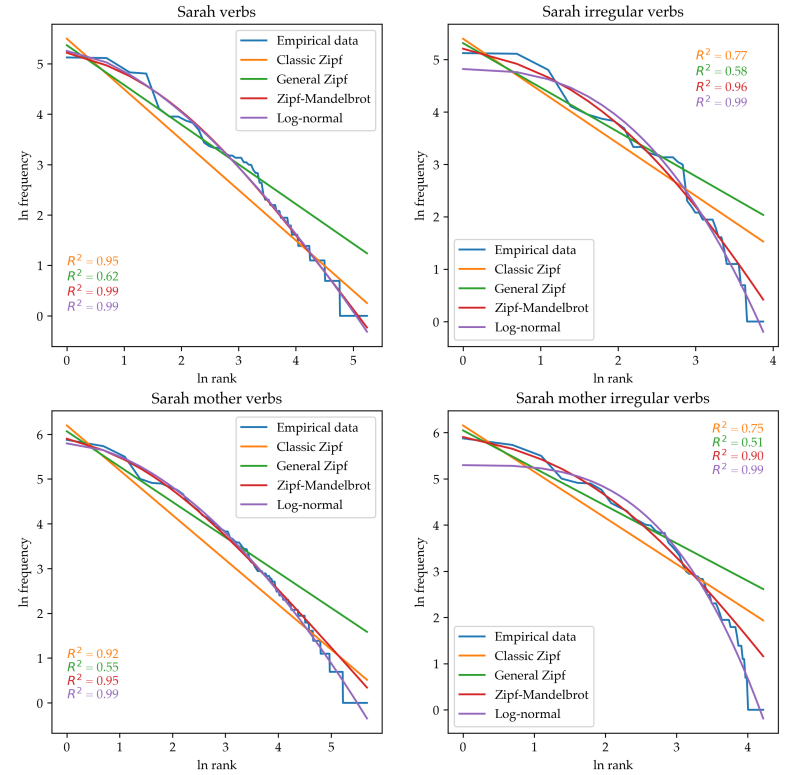

Eve verbs
$R^2 = 0.88$
$R^2 = 0.61$
$R^2 = 0.92$
$R^2 = 0.98$
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
ln frequency
ln rank

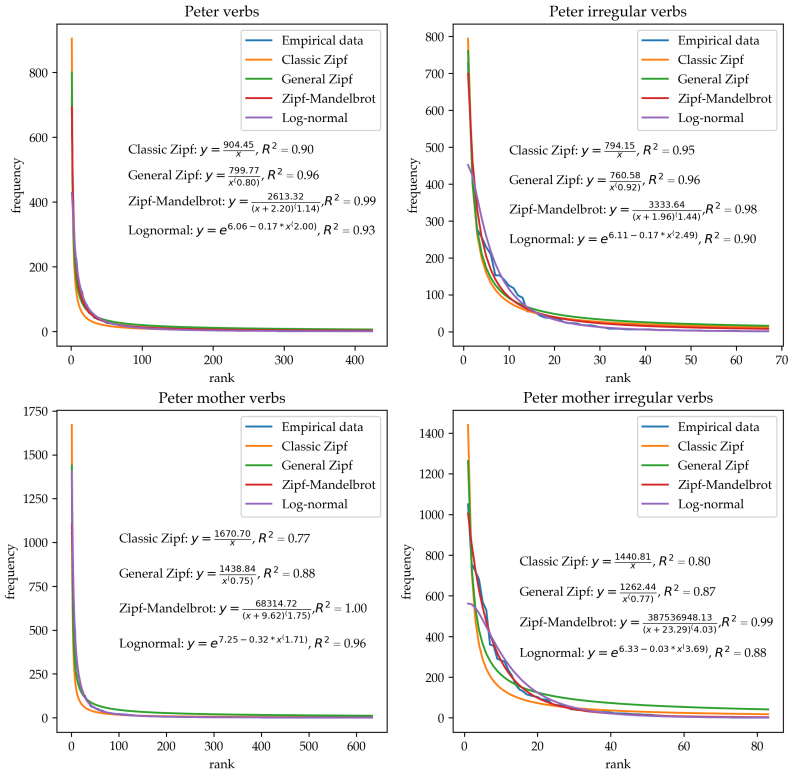
Eve irregular verbs
Empirical data
Classic Zipf
General Zipf
Log-normal
$R^2 = 0.85$
$R^2 = 0.53$
$R^2 = 0.96$
$R^2 = 0.98$
ln frequency
ln rank

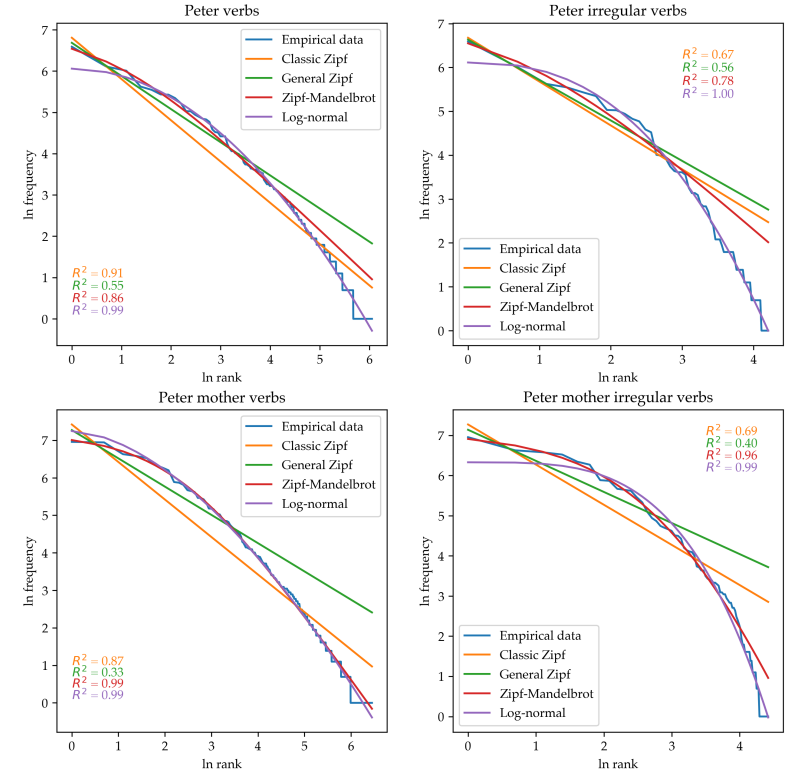

Eve mother verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
$R^2 = 0.92$
$R^2 = 0.58$
$R^2 = 0.98$
$R^2 = 0.99$
ln frequency
ln rank

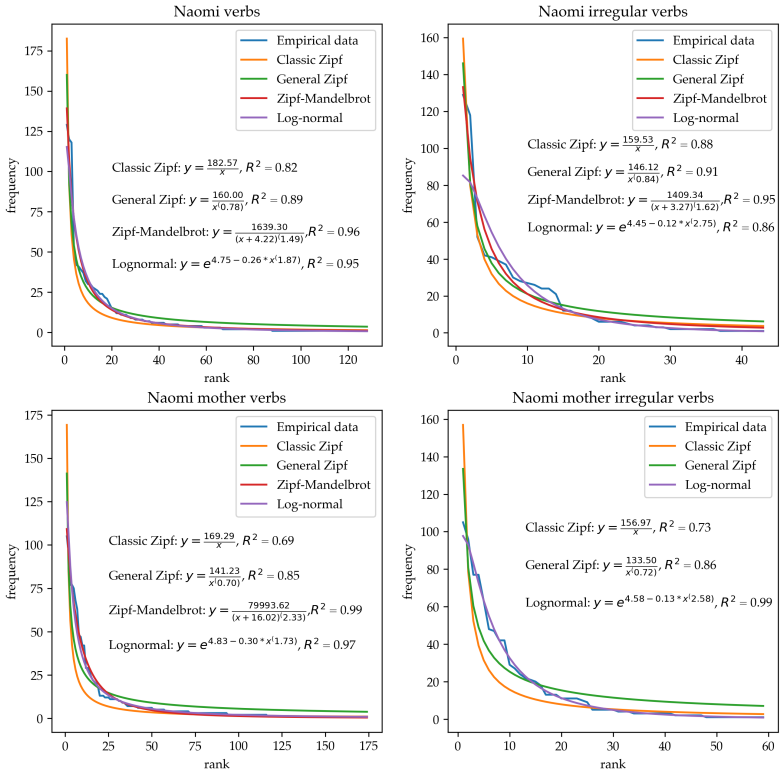
Eve mother irregular verbs
$R^2 = 0.81$
$R^2 = 0.60$
$R^2 = 0.95$
$R^2 = 0.99$
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
ln frequency
ln rank

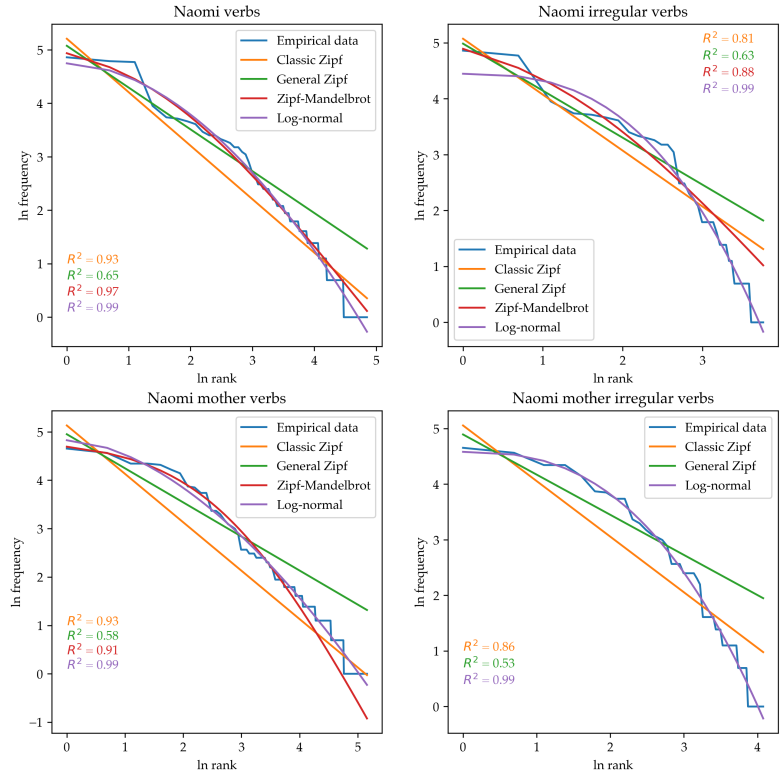

Sarah verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{243.50}{x}$, $R^2 = 0.83$
General Zipf: $y = \frac{214.00}{x^{(0.79)}}$, $R^2 = 0.90$
Zipf-Mandelbrot: $y = \frac{2743.74}{(x+4.73)^{(1.55)}}$, $R^2 = 0.97$
Lognormal: $y = e^{5.25 - 0.41 * x^{(1.58)}}$, $R^2 = 0.97$
frequency
rank
Sarah irregular verbs
Classic Zipf: $y = \frac{221.13}{x}$, $R^2 = 0.86$
General Zipf: $y = \frac{203.29}{x^{(0.85)}}$, $R^2 = 0.89$
Zipf-Mandelbrot: $y = \frac{36480.90}{(x+7.22)^{(2.51)}}$, $R^2 = 0.96$
Lognormal: $y = e^{4.82 - 0.15 * x^{(2.58)}}$, $R^2 = 0.90$
Sarah mother verbs
Classic Zipf: $y = \frac{490.63}{x}$, $R^2 = 0.86$
General Zipf: $y = \frac{430.65}{x^{(0.79)}}$, $R^2 = 0.93$
Zipf-Mandelbrot: $y = \frac{2949.60}{(x+3.74)^{(1.34)}}$, $R^2 = 0.99$
Lognormal: $y = e^{5.80 - 0.29 * x^{(1.76)}}$, $R^2 = 0.99$
Sarah mother irregular verbs
Classic Zipf: $y = \frac{470.87}{x}$, $R^2 = 0.87$
General Zipf: $y = \frac{423.16}{x^{(0.81)}}$, $R^2 = 0.92$
Zipf-Mandelbrot: $y = \frac{25998.37}{(x+6.69)^{(2.09)}}$, $R^2 = 0.99$
Lognormal: $y = e^{5.30 - 0.05 * x^{(3.25)}}$, $R^2 = 0.86$

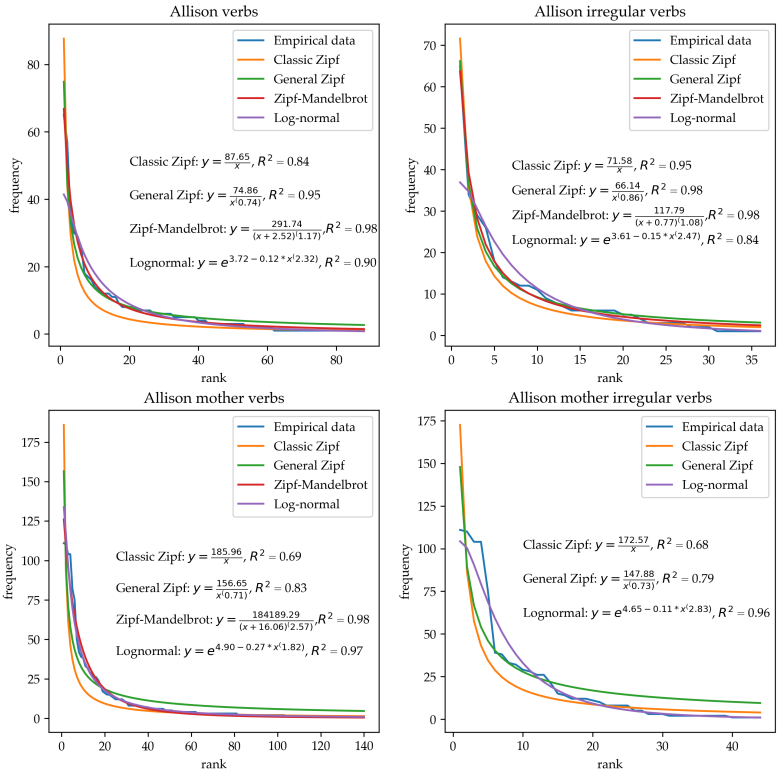

Sarah verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
$R^2 = 0.95$
$R^2 = 0.62$
$R^2 = 0.99$
$R^2 = 0.99$
ln frequency
ln rank
Sarah irregular verbs
$R^2 = 0.77$
$R^2 = 0.58$
$R^2 = 0.96$
$R^2 = 0.99$
Sarah mother verbs
$R^2 = 0.92$
$R^2 = 0.55$
$R^2 = 0.95$
$R^2 = 0.99$
Sarah mother irregular verbs
$R^2 = 0.75$
$R^2 = 0.51$
$R^2 = 0.90$
$R^2 = 0.99$

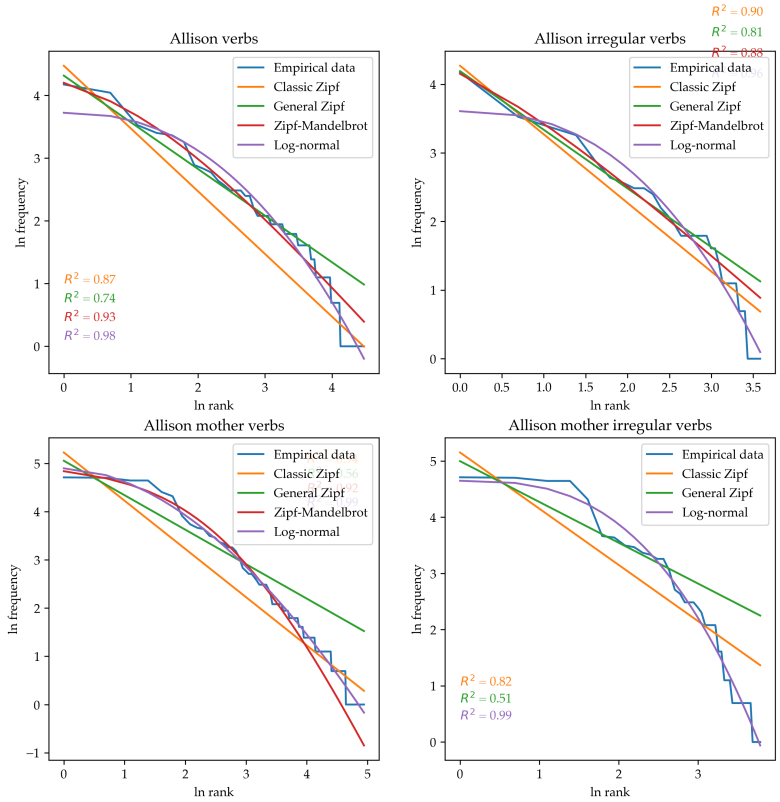

Peter verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{904.45}{x}$, $R^2 = 0.90$
General Zipf: $y = \frac{799.77}{x^{(0.80)}}$, $R^2 = 0.96$
Zipf-Mandelbrot: $y = \frac{2613.32}{(x+2.20)^{(1.14)}}$, $R^2 = 0.99$
Lognormal: $y = e^{6.06 - 0.17 * x^{(2.00)}}$, $R^2 = 0.93$
frequency
rank
Peter irregular verbs
Classic Zipf: $y = \frac{794.15}{x}$, $R^2 = 0.95$
General Zipf: $y = \frac{760.58}{x^{(0.92)}}$, $R^2 = 0.96$
Zipf-Mandelbrot: $y = \frac{3333.64}{(x+1.96)^{(1.44)}}$, $R^2 = 0.98$
Lognormal: $y = e^{6.11 - 0.17 * x^{(2.49)}}$, $R^2 = 0.90$
Peter mother verbs
Classic Zipf: $y = \frac{1670.70}{x}$, $R^2 = 0.77$
General Zipf: $y = \frac{1438.84}{x^{(0.75)}}$, $R^2 = 0.88$
Zipf-Mandelbrot: $y = \frac{68314.72}{(x+9.62)^{(1.75)}}$, $R^2 = 1.00$
Lognormal: $y = e^{7.25 - 0.32 * x^{(1.71)}}$, $R^2 = 0.96$
Peter mother irregular verbs
Classic Zipf: $y = \frac{1440.81}{x}$, $R^2 = 0.80$
General Zipf: $y = \frac{1262.44}{x^{(0.77)}}$, $R^2 = 0.87$
Zipf-Mandelbrot: $y = \frac{387536948.13}{(x+23.29)^{(4.03)}}$, $R^2 = 0.99$
Lognormal: $y = e^{6.33 - 0.03 * x^{(3.69)}}$, $R^2 = 0.88$

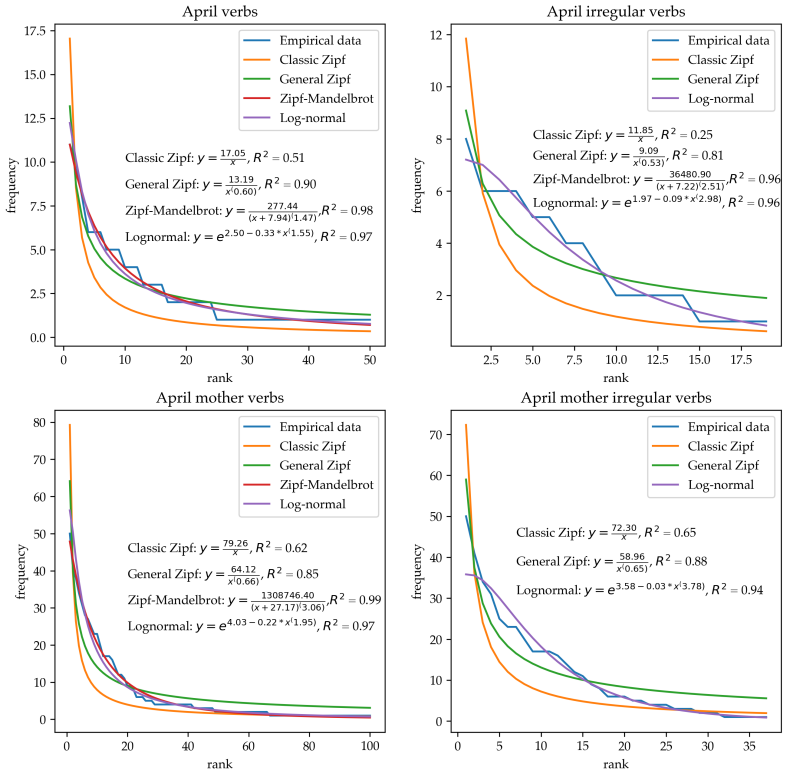

Peter verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
$R^2 = 0.91$
$R^2 = 0.55$
$R^2 = 0.86$
$R^2 = 0.99$
ln frequency
ln rank
Peter irregular verbs
$R^2 = 0.67$
$R^2 = 0.56$
$R^2 = 0.78$
$R^2 = 1.00$
Peter mother verbs
$R^2 = 0.87$
$R^2 = 0.33$
$R^2 = 0.99$
$R^2 = 0.99$
Peter mother irregular verbs
$R^2 = 0.69$
$R^2 = 0.40$
$R^2 = 0.96$
$R^2 = 0.99$

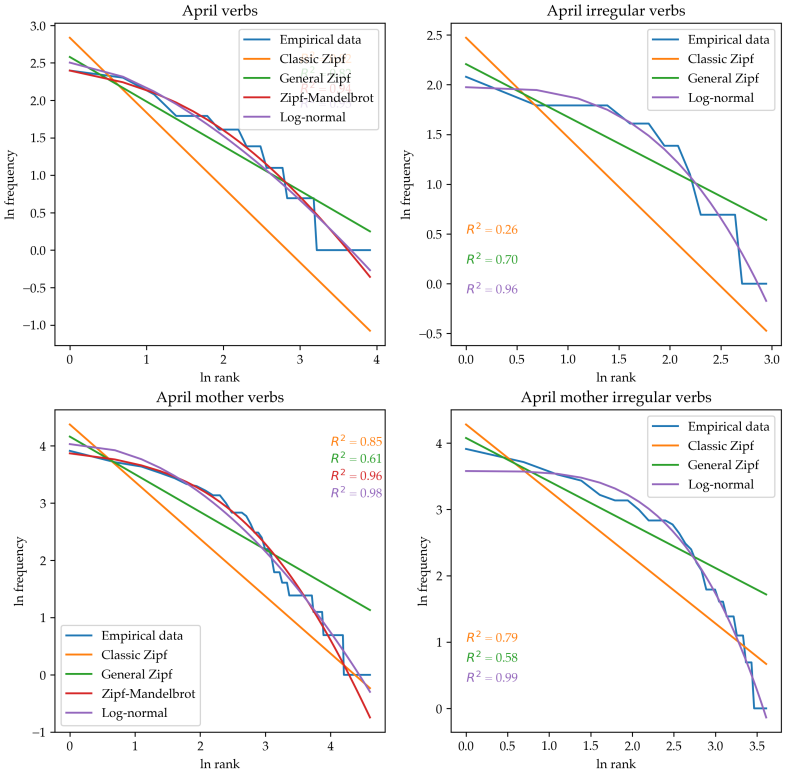

Naomi verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{182.57}{x}$, $R^2 = 0.82$
General Zipf: $y = \frac{160.00}{x^{(0.78)}}$, $R^2 = 0.89$
Zipf-Mandelbrot: $y = \frac{1639.30}{(x+4.22)^{(1.49)}}$, $R^2 = 0.96$
Lognormal: $y = e^{4.75 - 0.26 * x^{(1.87)}}$, $R^2 = 0.95$
frequency
rank
Naomi irregular verbs
Classic Zipf: $y = \frac{159.53}{x}$, $R^2 = 0.88$
General Zipf: $y = \frac{146.12}{x^{(0.84)}}$, $R^2 = 0.91$
Zipf-Mandelbrot: $y = \frac{1409.34}{(x+3.27)^{(1.62)}}$, $R^2 = 0.95$
Lognormal: $y = e^{4.45 - 0.12 * x^{(2.75)}}$, $R^2 = 0.86$
Naomi mother verbs
Classic Zipf: $y = \frac{169.29}{x}$, $R^2 = 0.69$
General Zipf: $y = \frac{141.23}{x^{(0.70)}}$, $R^2 = 0.85$
Zipf-Mandelbrot: $y = \frac{79993.62}{(x+16.02)^{(2.33)}}$, $R^2 = 0.99$
Lognormal: $y = e^{4.83 - 0.30 * x^{(1.73)}}$, $R^2 = 0.97$
Naomi mother irregular verbs
Classic Zipf: $y = \frac{156.97}{x}$, $R^2 = 0.73$
General Zipf: $y = \frac{133.50}{x^{(0.72)}}$, $R^2 = 0.86$
Lognormal: $y = e^{4.58 - 0.13 * x^{(2.58)}}$, $R^2 = 0.99$

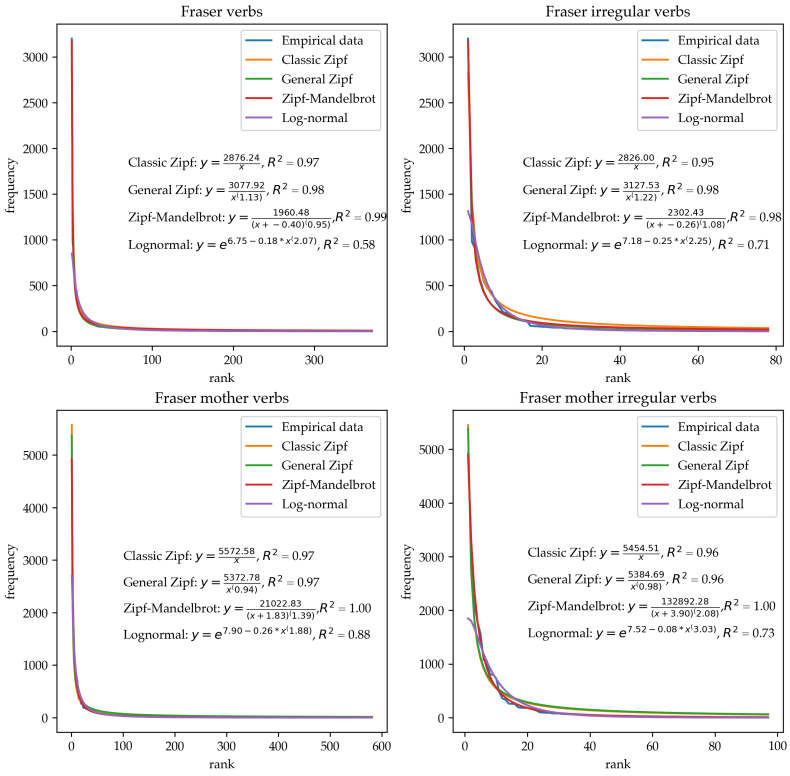

Naomi verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
$R^2 = 0.93$
$R^2 = 0.65$
$R^2 = 0.97$
$R^2 = 0.99$
ln frequency
ln rank
Naomi irregular verbs
$R^2 = 0.81$
$R^2 = 0.63$
$R^2 = 0.88$
$R^2 = 0.99$
Naomi mother verbs
$R^2 = 0.93$
$R^2 = 0.58$
$R^2 = 0.91$
$R^2 = 0.99$
Naomi mother irregular verbs
$R^2 = 0.86$
$R^2 = 0.53$
$R^2 = 0.99$

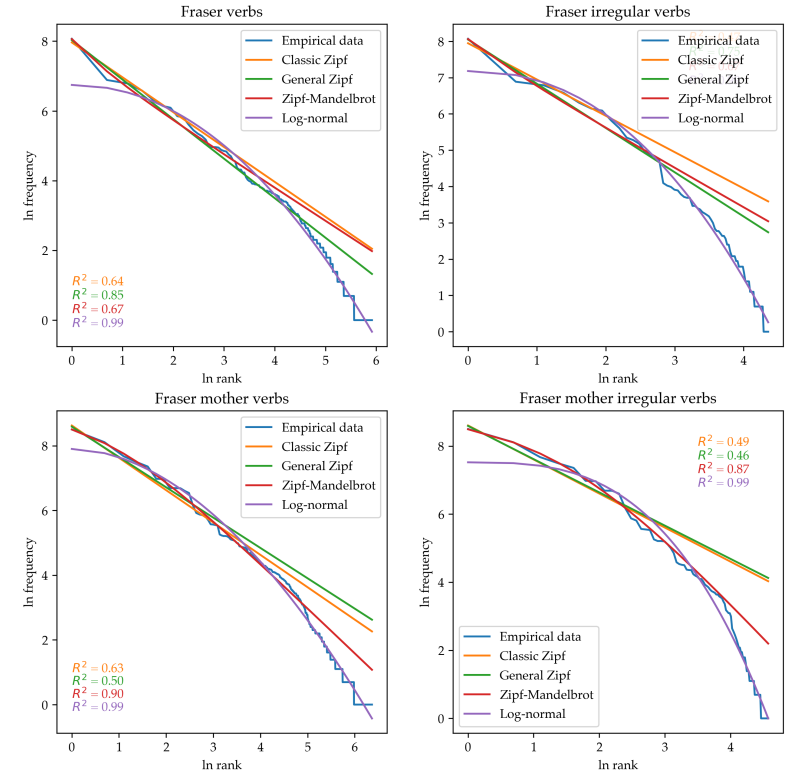

Allison verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{87.65}{x}$, $R^2 = 0.84$
General Zipf: $y = \frac{74.86}{x^{(0.74)}}$, $R^2 = 0.95$
Zipf-Mandelbrot: $y = \frac{291.74}{(x+2.52)^{(1.17)}}$, $R^2 = 0.98$
Lognormal: $y = e^{3.72 - 0.12 * x^{(2.32)}}$, $R^2 = 0.90$
frequency
rank
Allison irregular verbs
Classic Zipf: $y = \frac{71.58}{x}$, $R^2 = 0.95$
General Zipf: $y = \frac{66.14}{x^{(0.86)}}$, $R^2 = 0.98$
Zipf-Mandelbrot: $y = \frac{117.79}{(x+0.77)^{(1.08)}}$, $R^2 = 0.98$
Lognormal: $y = e^{3.61 - 0.15 * x^{(2.47)}}$, $R^2 = 0.84$
Allison mother verbs
Classic Zipf: $y = \frac{185.96}{x}$, $R^2 = 0.69$
General Zipf: $y = \frac{156.65}{x^{(0.71)}}$, $R^2 = 0.83$
Zipf-Mandelbrot: $y = \frac{184189.29}{(x+16.06)^{(2.57)}}$, $R^2 = 0.98$
Lognormal: $y = e^{4.90 - 0.27 * x^{(1.82)}}$, $R^2 = 0.97$
Allison mother irregular verbs
Classic Zipf: $y = \frac{172.57}{x}$, $R^2 = 0.68$
General Zipf: $y = \frac{147.88}{x^{(0.73)}}$, $R^2 = 0.79$
Lognormal: $y = e^{4.65 - 0.11 * x^{(2.83)}}$, $R^2 = 0.96$


Allison verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
$R^2 = 0.87$
$R^2 = 0.74$
$R^2 = 0.93$
$R^2 = 0.98$
ln frequency
ln rank
Allison irregular verbs
$R^2 = 0.90$
$R^2 = 0.81$
$R^2 = 0.88$
Allison mother verbs
Allison mother irregular verbs
$R^2 = 0.82$
$R^2 = 0.51$
$R^2 = 0.99$


April verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{17.05}{x}$, $R^2 = 0.51$
General Zipf: $y = \frac{13.19}{x^{0.60}}$, $R^2 = 0.90$
Zipf-Mandelbrot: $y = \frac{277.44}{(x+7.94)^{1.47}}$, $R^2 = 0.98$
Lognormal: $y = e^{2.50 - 0.33 * x^{1.55}}$, $R^2 = 0.97$
frequency
rank
April irregular verbs
Classic Zipf: $y = \frac{11.85}{x}$, $R^2 = 0.25$
General Zipf: $y = \frac{9.09}{x^{0.53}}$, $R^2 = 0.81$
Zipf-Mandelbrot: $y = \frac{36480.90}{(x+7.22)^{2.51}}$, $R^2 = 0.96$
Lognormal: $y = e^{1.97 - 0.09 * x^{2.98}}$, $R^2 = 0.96$
April mother verbs
Classic Zipf: $y = \frac{79.26}{x}$, $R^2 = 0.62$
General Zipf: $y = \frac{64.12}{x^{0.66}}$, $R^2 = 0.85$
Zipf-Mandelbrot: $y = \frac{1308746.40}{(x+27.17)^{3.06}}$, $R^2 = 0.99$
Lognormal: $y = e^{4.03 - 0.22 * x^{1.95}}$, $R^2 = 0.97$
April mother irregular verbs
Classic Zipf: $y = \frac{72.30}{x}$, $R^2 = 0.65$
General Zipf: $y = \frac{58.96}{x^{0.65}}$, $R^2 = 0.88$
Lognormal: $y = e^{3.58 - 0.03 * x^{3.78}}$, $R^2 = 0.94$


April verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
ln frequency
ln rank
April irregular verbs
$R^2 = 0.26$
$R^2 = 0.70$
$R^2 = 0.96$
April mother verbs
$R^2 = 0.85$
$R^2 = 0.61$
$R^2 = 0.96$
$R^2 = 0.98$
April mother irregular verbs
$R^2 = 0.79$
$R^2 = 0.58$
$R^2 = 0.99$


Fraser verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
Classic Zipf: $y = \frac{2876.24}{x}$, $R^2 = 0.97$
General Zipf: $y = \frac{3077.92}{x^{(}1.13)}$, $R^2 = 0.98$
Zipf-Mandelbrot: $y = \frac{1960.48}{(x + -0.40)^{(}0.95)}$, $R^2 = 0.99$
Lognormal: $y = e^{6.75 - 0.18 * x^{(}2.07)}$, $R^2 = 0.58$
frequency
rank
Fraser irregular verbs
Classic Zipf: $y = \frac{2826.00}{x}$, $R^2 = 0.95$
General Zipf: $y = \frac{3127.53}{x^{(}1.22)}$, $R^2 = 0.98$
Zipf-Mandelbrot: $y = \frac{2302.43}{(x + -0.26)^{(}1.08)}$, $R^2 = 0.98$
Lognormal: $y = e^{7.18 - 0.25 * x^{(}2.25)}$, $R^2 = 0.71$
Fraser mother verbs
Classic Zipf: $y = \frac{5572.58}{x}$, $R^2 = 0.97$
General Zipf: $y = \frac{5372.78}{x^{(}0.94)}$, $R^2 = 0.97$
Zipf-Mandelbrot: $y = \frac{21022.83}{(x + 1.83)^{(}1.39)}$, $R^2 = 1.00$
Lognormal: $y = e^{7.90 - 0.26 * x^{(}1.88)}$, $R^2 = 0.88$
Fraser mother irregular verbs
Classic Zipf: $y = \frac{5454.51}{x}$, $R^2 = 0.96$
General Zipf: $y = \frac{5384.69}{x^{(}0.98)}$, $R^2 = 0.96$
Zipf-Mandelbrot: $y = \frac{132892.28}{(x + 3.90)^{(}2.08)}$, $R^2 = 1.00$
Lognormal: $y = e^{7.52 - 0.08 * x^{(}3.03)}$, $R^2 = 0.73$


Fraser verbs
Empirical data
Classic Zipf
General Zipf
Zipf-Mandelbrot
Log-normal
$R^2 = 0.64$
$R^2 = 0.85$
$R^2 = 0.67$
$R^2 = 0.99$
ln frequency
ln rank
Fraser irregular verbs
Fraser mother verbs
$R^2 = 0.63$
$R^2 = 0.50$
$R^2 = 0.90$
$R^2 = 0.99$
Fraser mother irregular verbs
$R^2 = 0.49$
$R^2 = 0.46$
$R^2 = 0.87$
$R^2 = 0.99$

References

Lada A Adamic and Bernardo A Huberman. Zipf's law and the internet. *Glottometrics*, 3(1): 143–150, 2002.

Jeff Alstott and Dietmar Plenz Bullmore. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1), 2014.

Stephen Robert Anderson. *West Scandinavian vowel systems and the ordering of phonological rules.* PhD thesis, Massachusetts Institute of Technology, 1969.

R Harald Baayen. *Word frequency distributions*, volume 18. Springer Science & Business Media, 2002.

Curtis A Becker. Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 5(2):252, 1979.

Jean Berko. The child's learning of english morphology. *Word*, 14(2-3):150–177, 1958.

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.

JB Carrol. On sampling from a lognormal model of word frequency distribution. *Computational analysis of present-day American English*, pages 406–424, 1967.

John B Carroll. A rationale for an asymptotic lognormal form of word-frequency distributions. *ETS Research Bulletin Series*, 1969(2):i–94, 1969.

Andrew W Ellis and Catriona M Morrison. Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(2):515, 1998.

Kenneth I Forster. Accessing the mental lexicon. *New approaches to language mechanisms*, pages 257–287, 1976.

Kenneth I Forster. Memory-addressing mechanisms and lexical access. *Orthography, phonology, morphology, and meaning*, page 413, 1992.

Simon Gerhand and Christopher Barry. Age-of-acquisition and frequency effects in speeded word naming. *Cognition*, 73(2):B27–B36, 1999.

Morris Halle and Alec Marantz. Distributed morphology and the pieces of inflection. hale, k. & sj keyser (eds.), the view from building 20, 1993.

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

Davis H Howes and Richard L Solomon. Visual duration threshold as a function of word-probability. *Journal of experimental psychology*, 41(6):401, 1951.

Paul Kiparsky. *" Elsewhere" in phonology.* Indiana University Linguistics Club, 1973.

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. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press, 2000.

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

Benoît Mandelbrot. Information theory and psycholinguistics. *BB Wolman and E*, 1965.

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.

LM McCusker. Some determinants of word recognition: Frequency. In *24th annual convention of the southwestern psychological association, fort worth, tx*, 1977.

Isabel Moreno-Sánchez, Francesc Font-Clos, and Álvaro Corral. Large-scale analysis of zipf's law in english texts. *PloS one*, 11(1), 2016.

Catriona M Morrison and Andrew W Ellis. Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1):116, 1995.

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

Kenneth R Paap, Sandra L Newsome, James E McDonald, and Roger W Schvaneveldt. An activation–verification model for letter and word recognition: The word-superiority effect. *Psychological review*, 89(5):573, 1982.

Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

Steven Pinker and Michael T Ullman. The past and future of the past tense. *Trends in cognitive sciences*, 6(11):456–463, 2002.

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 P Whaley. Word—nonword classification time. *Journal of Verbal learning and Verbal behavior*, 17(2):143–154, 1978.

Charles Yang. Dig-dug, think-thunk. *The London Review of Books*, 22(10), 2000.

Charles Yang. On productivity. *Linguistic variation yearbook*, 5(1):265–302, 2005.

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

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

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

George Kingsley Zipf. The psycho-biology of language.(1935). *List of Figures List of Figures List of Figures*, 1935.

George Kingsley Zipf. Human behavior and the principle of least effort. 1949.