Testing the Tolerance Principle

Xiaomeng Ma

Tolerance Principle

Definition

Rule-based learning, such as past-tense acquisition, is commonly observed in language acquisition. But what leads to the use of the rules in the first place?

> **Tolerance Principle**: Let *R* be a rule that applies to N items, of which e are exceptions. *R* is productive if and only iff: $e \leq \theta_N$, where $\theta = \dfrac{N}{lnN}$ (Yang, 2016).

Examples:

- If we have 20 verbs, and 5 are irregular verbs, will a rule be derived?
  - N = 20, e = 5, θ = $\dfrac{N}{lnN} = \dfrac{20}{ln(20)} \approx$ 6.7.
  - e $< \theta$, so a rule **will** be derived
- If we have 10 verbs, and 5 are irregular verbs, will a rule be derived?
  - N = 10, e = 5, θ = $\dfrac{N}{lnN} = \dfrac{10}{ln(10)} \approx$ 4.3.
  - e $> \theta$, so a rule **won't** be derived

Assumptions of the Tolerance Principle

- Why a Rule is deployed? A productive rule should be deployed when it delivers more efficient results than not using the rule.
    - For TP, more efficient = Faster
    - TP hypothesizes that a productive rule will reduce the average time of retrieving the target form.
- Dual-route model for the regular and irregular verb processing (Pinker & Prince, 1988).

Dual-route Model

- Regular Verbs: Processed by Rule Applying Mechanism
- Irregular Verbs: Processed through rote and associative memory
- Input goes into LEXICON for search, the LEXICON only contains the suffix and the irregular forms. If a match is found, then output the irregular form; else, apply the rule.

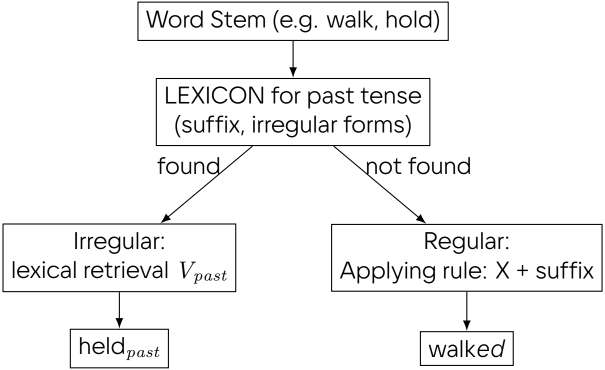
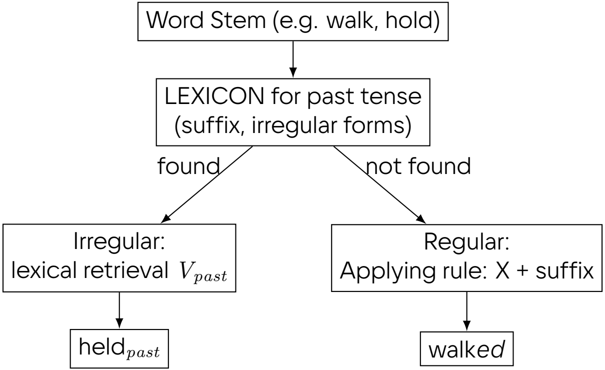


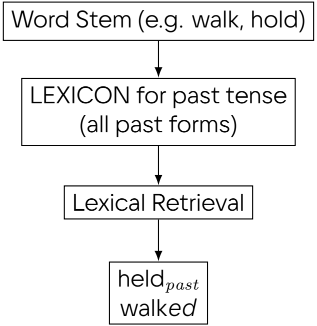
Table of Contents

Rule VS No Rule

Rule-based Model



No-Rule Model



Intuition: Rule-based Model should save lexical retrieval time than No-Rule Model because there are fewer items to search in the LEXICON.

No Rule: Calculating the time complexity (T)

- **LEXICON structure:** All the lexical items are stored in a ranked list based on their frequency, with the most frequent items at the top.
- **Search function:** Serial Search process (Forster, 1976, 1992): to retrieve an item at position i, the model sequentially searches all the i-1 items ranked higher than i.
- **Intuition:** the less frequent words take longer to retrieve than the more frequent words.
- **Each word's average time complexity:** $p_i \cdot t_i$, p_i is the word's probability and t_i is its retrieval time.
- **Average time complexity for the lexicon list:** $T = \sum_{i=1}^{N} (p_i \cdot t_i)$.

No Rule: Calculating the time complexity (T)

- t_i: Assumes the rank hypothesis (Murray & Forster, 2004): the i-th ranked item takes i units of time to be retrieved, $t_i = r_i$
- p_i: Assumes the Zipfian distribution (Zipf, 1949): a word's frequency (f_i) times its rank (r_i) is a constant C: $C = f_i \cdot r_i$.
- Replacing f_i with $\frac{C}{r_i}$, $p_i = \frac{f_i}{\sum_{k=1}^{N} f_k} = \frac{\frac{C}{r_i}}{\sum_{k=1}^{N} \frac{C}{r_k}} = \frac{\frac{1}{r_i}}{\sum_{k=1}^{N} \frac{1}{r_k}}$.
- Insert: $T = \sum_{i=1}^{N} (p_i \cdot t_i) = \sum_{i=1}^{N} (\frac{\frac{1}{r_i}}{\sum_{k=1}^{N} \frac{1}{r_k}} \cdot r_i) = \sum_{i=1}^{N} (\frac{1}{\sum_{k=1}^{N} \frac{1}{r_k}})$
- $\frac{1}{\sum_{k=1}^{N} \frac{1}{r_k}}$ is Harmonic number H_N and Yang approximated $H_N \approx lnN$
- $T_{NoRule} \approx \frac{N}{lnN}$

Rule: Calculating the time complexity

- Rule-based model divides the time complexity into two parts: T_E for the exceptions and T_R for the rule-based items.
- Assuming there are N items and e exceptions ($e \leq N$).
- **Exceptions** are processed the same way in the no-rule model: $T_E \approx \frac{e}{lne} \cdot \frac{e}{N}$
- **Rule-based items** are assumed to have the same time complexity because they are reached after a thorough search of e exceptions: $T_R = e \cdot (1 - \frac{e}{N})$.
- $T_{Rule} = \frac{e}{lne} \cdot \frac{e}{N} + e \cdot (1 - \frac{e}{N})$

Deriving the Tolerance Principle

Assumption: more efficient = faster, so when $T_{Rule} \leq T_{NoRule}$ a rule will be deployed. Solving the inequation:

$$\frac{e}{N} \cdot \frac{e}{lne} + (1 - \frac{e}{N}) \cdot e \leq \frac{N}{lnN} \tag{1}$$

$$\frac{e}{N} \cdot (\frac{e}{lne} - e) + e \leq \frac{N}{lnN} \tag{2}$$

Since $\frac{e}{N} \cdot (\frac{e}{lne} - e) \leq 0$, therefore $e \leq \frac{N}{lnN}$.

When $e \leq \frac{N}{lnN}$, a rule will be deployed.

Table of Contents

Problems with estimation $\frac{N}{lnN}$

Mathematically: $\lim_{x \to \infty} (H_N - lnN) = \gamma$ where γ is Euler's constant $\approx$ 0.58

- The difference between H_N and lnN could be substantial for TP's calculation
- For example, when N = 10: $\frac{N}{lnN} \approx 4.34$, $\frac{N}{H_N} \approx 3.41$.
- When there are 4 exceptions:
  - lnN says **yes** can be a rule $(4 < 4.34)$.
  - H_N says **no** there can't be a rule $(4 > 3.41)$.

Problems with approximation of the inequation

Mathematically: $e \leq \frac{N}{lnN}$ is not the solution to $\frac{e}{N} \cdot (\frac{e}{lne} - e) + e \leq \frac{N}{lnN}$

- The difference between the **actual** solution of e and **estimated** $\frac{N}{lnN}$ could be substantial.
- For example, when N = 20: $\frac{N}{lnN} \approx 6.67$, therefore $e \leq 6.67$.
- However, the true solution to the inequation is $e \leq 8.73$.
- When there are 7 exceptions, can a rule be derived?

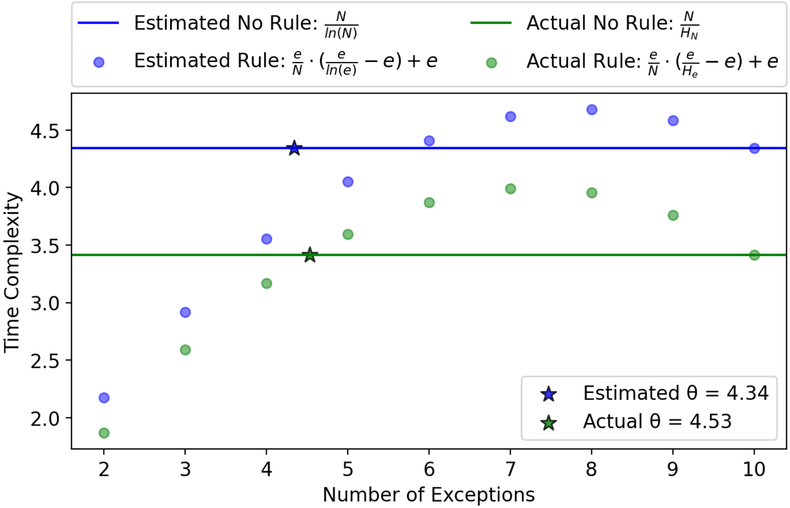
Testing it with different Ns

- Calculating the **actual** threshold θ using the Harmonic number by solving inequation:
$\frac{e}{N} \cdot \frac{e}{lne} + (1 - \frac{e}{N}) \cdot e \leq \frac{N}{lnN}$
- Comparing the result with $\frac{N}{lnN}$
- N = 10, 100, 1000

Table: The predicted θ and actual θ with different N

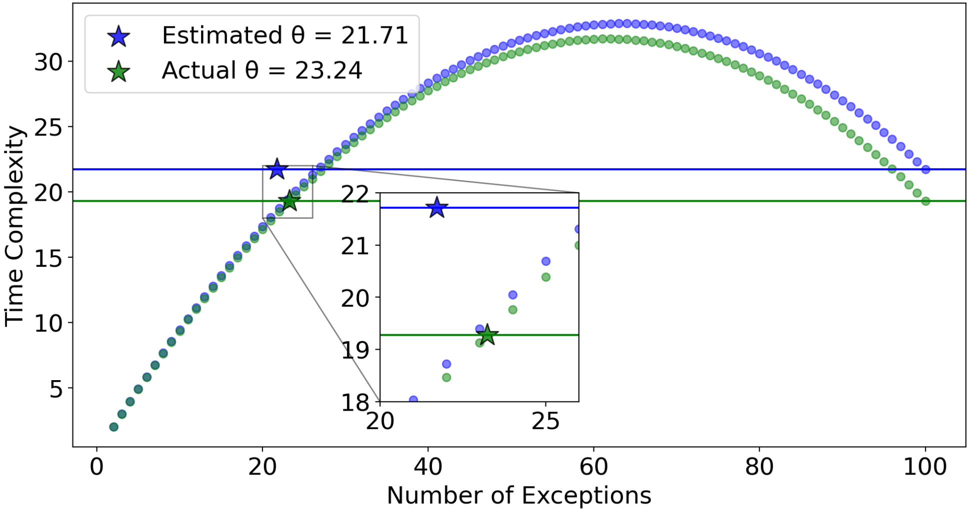
N	Predicted θ $N/ln(N)$	Actual θ
10	4.34	4.53
100	21.71	23.24
1,000	144.76	152.77

Plot: N = 10



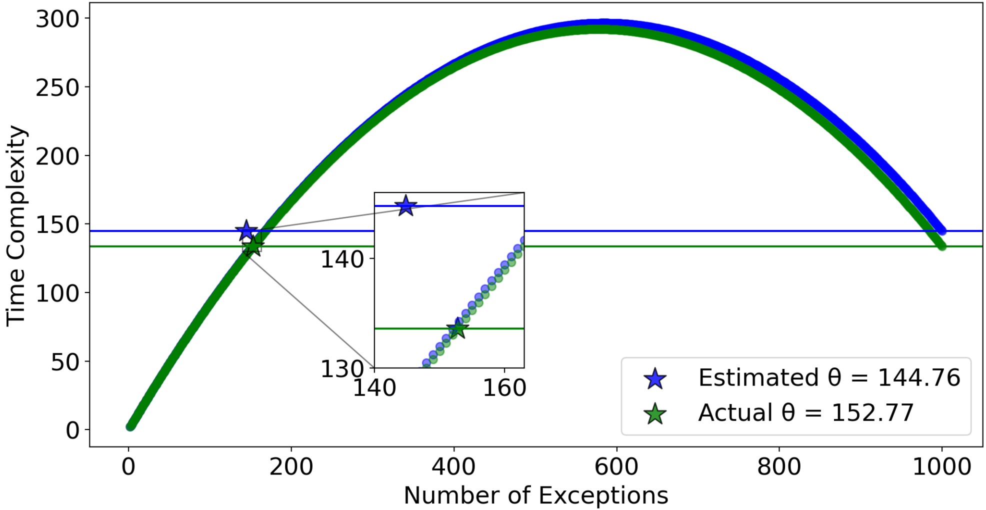
Estimated No Rule: $\frac{N}{ln(N)}$
Actual No Rule: $\frac{N}{H_N}$
Estimated Rule: $\frac{e}{N} \cdot (\frac{e}{ln(e)} - e) + e$
Actual Rule: $\frac{e}{N} \cdot (\frac{e}{H_e} - e) + e$
Time Complexity
4.5
4.0
3.5
3.0
2.5
2.0
2
3
4
5
6
7
8
9
10
Number of Exceptions
Estimated θ = 4.34
Actual θ = 4.53

Plot: N = 100



Estimated θ = 21.71
Actual θ = 23.24
Time Complexity
Number of Exceptions
0
20
40
60
80
100
5
10
15
20
25
30
22
21
20
19
18
20
25

Plot: N = 1000



300
250
200
150
100
50
0
Time Complexity
0
200
400
600
800
1000
Number of Exceptions
140
130
140
160
Estimated θ = 144.76
Actual θ = 152.77

Rank Matters

- **Intuition:** In TP's calculation, the t_i is the rank of the item. What if the exceptions all have high ranks? How would that affect the solution of the inequation?
- Creating a hypothetical list of 10 items whose distribution follows a Zipfian distribution: 1st ranked item has a frequency of 100 and the 10th ranked item has a frequency of 1.
- Calculate the T_{Rule} and T_{NoRule} using the formula and find the solution to the inequation $T_{Rule} \leq T_{NoRule}$.

Base Scenario: 10 items, All Exceptions

Intuition: No Rule should be derived

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

Scenario 1: 10 items, 7 exceptions

Intuition: No Rule should be derived.

No Rule

Verb	Frequency	rank	Time Complexity
Excep.	100	1	0.34
Excep.	50	2	0.34
Excep.	33	3	0.34
Excep.	25	4	0.34
Excep.	20	5	0.34
Excep.	17	6	0.35
Excep.	14	7	0.33
Regular	13	8	0.35
Regular	11	9	0.34
Regular	10	10	0.34
Total	293	T_{NoRule}	3.42

With a Rule

Excep.	Freq.	rank	Time Complexity
Excep.	100	1	0.24 = 100/293 x 1 x 0.7
Excep.	50	2	0.24 = 50/293 x 2 x 0.7
Excep.	33	3	0.24 = 33/293 x 3 x 0.7
Excep.	25	4	0.24 = 25/293 x 4 x 0.7
Excep.	20	5	0.24 = 20/293 x 5 x 0.7
Excep.	17	6	0.24 = 17/293 x 6 x 0.7
Excep.	14	7	0.23 = 14/293 x 7 x 0.7
Total			**1.67**
Regular			Time
Regular	13		
Regular	11		
Regular	10		
Total			**2.1** = 7 x 0.3
T_{Rule}			**3.77** = 1.67 + 2.1 $>$ 3.42

$T_{Rule} = 3.77$, $T_{NoRule} = 3.42$, since $T_{Rule} > T_{NoRule}$, **No rule will be derived.**

Scenario 2: 10 items, 7 exceptions

Intuition: No Rule should be derived.

No Rule

Item	Frequency	rank	Time Complexity
Excep.	100	1	0.34
Regular	50	2	0.34
Excep.	33	3	0.34
Excep.	25	4	0.34
Regular	20	5	0.34
Regular	17	6	0.35
Excep.	14	7	0.33
Excep.	13	8	0.35
Excep.	11	9	0.34
Excep.	10	10	0.34
Total	293	T_{NoRule}	3.42

With a Rule

Excep.	Freq.	rank	Time Complexity $T_E = \sum_{i=1}^{N}(p_i \cdot t_i) \cdot \frac{e}{N}$
Excep	100	1	0.24 = 100/293 x 1 x 0.7
Excep	33	2	0.16 = 33/293 x 2 x 0.7
Excep	25	3	0.18 = 25/293 x 3 x 0.7
Excep	14	4	0.13 = 14/293 x 4 x 0.7
Excep	13	5	0.16 = 13/293 x 5 x 0.7
Excep	11	6	0.16 = 11/293 x 6 x 0.7
Excep	10	7	0.17 = 10/293 x 7 x 0.7
Total			**1.19**

Regular			Time Complexity
Regular	50		$T_R = e \cdot (1 - \frac{e}{N})$
Regular	20		
Regular	17		
Total			**2.1** = 7 x 0.3
T_{Rule}			**3.29** = 1.19 + 2.1 $<$ 3.42

$T_{Rule} = 3.29$, $T_{NoRule} = 3.42$, since $T_{Rule} < T_{NoRule}$, **rule will be derived.**

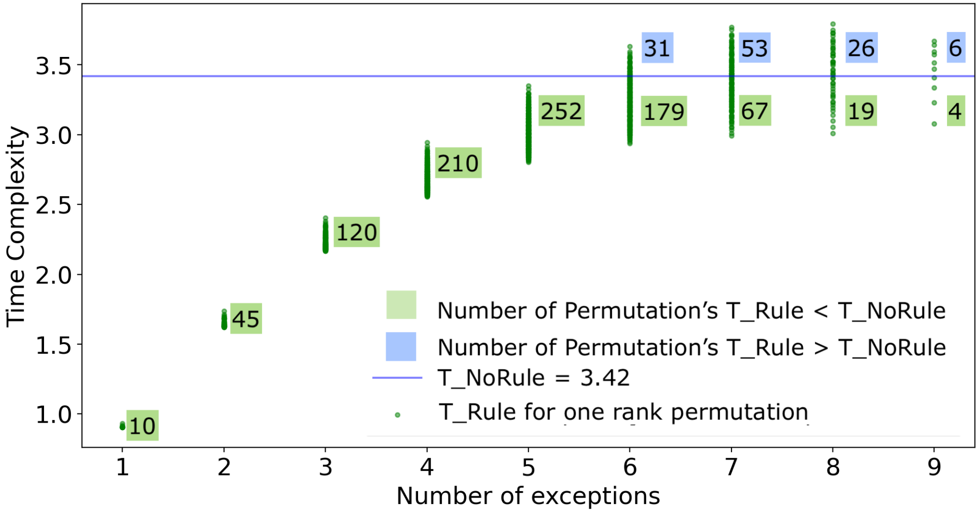
Discrepancies of the TP

- According to the TP, the number of exceptions (e) is the **only** factor determining whether a rule will be derived.
- Depending on the rank of the exceptions, the same number of exceptions would produce contradicting results. (e.g. In Scenario 1, $T_{Rule} < T_{NoRule}$, a rule will be derived. In Scenario 2, $T_{Rule} > T_{NoRule}$, a rule won't be derived.)
- Time complexity is not a fixed value. It varies depending on the rank permutation.

Testing the Rank Permutation

- **Observation:** When the regulars are of highest ranks, T_{Rule} reaches its **maximum**. When the regulars are of the lowest ranks, T_{Rule} reaches its **minimum**.
- Using N = 10 exhaustively calculate the T_{Rule} for all rank permutations with different numbers of exceptions
- Solve the inequation $T_{Rule} < T_{NoRule}$ to find the threshold θ.

Permuation: N = 10

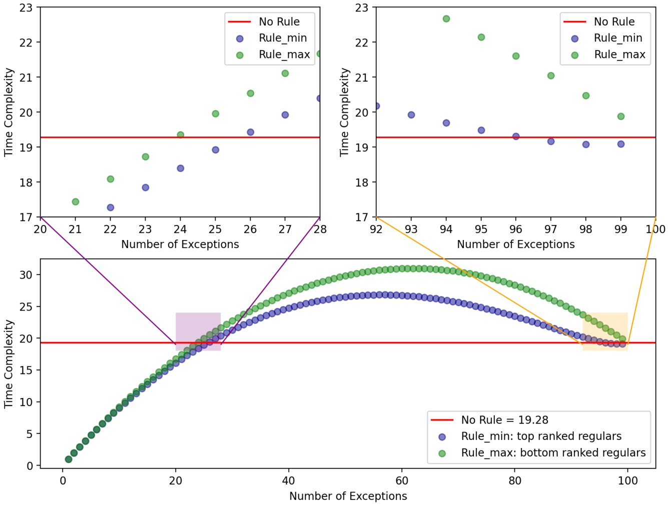


Further Test: N = 100, 1000

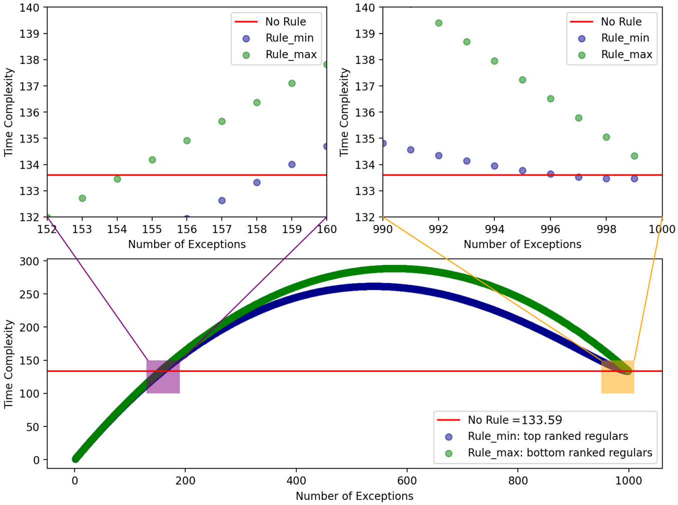
- Using N = 100, 1000 to calculate the $T_{Rule(MAX)}$ and $T_{Rule(MIN)}$ and find θ_{min} and θ_{max}.

	N = 100	N = 1000
T_{NoRule}	19.28	133.59
$\theta = N/ln(N)$	21.71	144.76
$T_{Rule(MAX)}$ (with integer θ)		
θ_{min}	23	154
θ_{max}	NA(> 100)	NA(> 1000)
$T_{Rule(MIN)}$ (with integer θ)		
θ_{min}	25	158
θ_{max}	97	997
A rule is derived when $e \leq \theta_{min}$ or $e \geq \theta_{max}$		

Permuation: N = 100



Permuation: N = 1000



Summary

Mathematical Discrepancies

- $\frac{N}{lnN}$ is not a proper estimation of the maximum number of exceptions.
- $e \leq \frac{N}{lnN}$ is not a proper solution to the inequation $T_{Rule} < T_{NoRule}$

When a rule is derived cannot be solely predicted on e

- Datasets with the same number of exceptions but different rank permutations can lead to contradicting results.

More Problems

Quadratic function of T_{Rule}

- TP's assumption assumes that the exceptions and the T_{Rule} has a linear relationship: if e is smaller than a threshold, $T_{Rule} < T_{NoRule}$, thus a rule will be derived.
- However, T_{Rule} is obviously quadratic, there are two sets of data that fit the rule-deriving criterion $T_{Rule} < T_{NoRule}$: $e \leq \theta_{min}$ or $e \geq \theta_{max}$.
- For example, when N = 100 and 1000, and $e > 97$ or $e > 997$, $T_{Rule} < T_{NoRule}$, a rule can be derived, which is impossible.
- The basic assumption of the TP is flawed.

Table of Contents

Motivation

One may argue that in real life many of the frequency permutations are not plausible and the quadratic pattern doesn't apply since there is a fixed number of exceptions.

Test the TP on past tense overregularization using children's corpus data.

Yang (2016)'s testing

- Yang (2016) applied the TP to explain past tense acquisition on Adam's and Eve's data (Brown, 1973; MacWhinney, 2000).
- The first overregularization error (e.g. **holded*) is seen as the sign or rule being deployed.
- Data: first recording to the recording of overregularization error
  - Adam: 2;3 - 2;11
  - Eve: 1:6 - 1:10*
- N: all the verb forms (including -*ing*, verb root, etc)the child produced.
- e: all the irregular verb forms the child produced.
- Results
  - Adam: N = 300, e = 57, $\theta = N/lnN \approx 53$, 57 $>$ 53, **failed** ✕
  - Eve: N = 163, e = 49, $\theta \approx 32$, **failed** ✕
- Explanation: Sampling errors

New Testing: Data

8 children's data from CHILDES.

	Age range	files	Total Verb Types (N)	Irregular Types (e)	Total Verb tokens	Irregular tokens
Adam	2;3 - 2;11	18	306	62	6,747	3,632
Eve[1]	1;6 - 1;8	5	93	36	564	337
Sarah	2;3 - 2;10	33	189	48	1,759	1,035
Peter	1;3 - 2;6	14	424	67	7,532	3,647
Naomi	1;3 - 1;11	20	128	43	1,240	757
Allison	1;5 - 2;11	6	88	36	612	335
April	1;10 - 2;1	2	50	19	128	62
Fraser	2;0 - 2;5	90	371	78	13,924	9,903

New Testing: Method

- Replicated Yang's method to count N and e.
- Compare e to N/lnN
- Compare e to $N/H(N)$
- Calculate T_{Rule} and T_{NoRule} using the verbs' actual rank and frequency and compare

New Testing: Results

	N	e	θ_p	$e < \theta_p$	θ_a	$e < \theta_a$	T_{NoRule}	T_{Rule}	$T_{Rule} < T_{NoRule}$
Adam	306	62	53.5	×	48.6	×	33.80	51.33	×
Eve	93	36	20.5	×	18.2	×	17.51	25.11	×
Sarah	189	48	36.1	×	32.5	×	25.65	37.81	×
Peter	424	67	70.1	✓	64	×	43.82	57.74	×
Naomi	128	43	26.4	×	23.6	×	19.63	31.23	×
Allison	88	36	19.7	×	17.4	×	18.24	34.72	×
April	50	19	12.8	×	11.1	×	14.64	14.29	✓
Fraser	371	78	62.7	×	57.1	×	26.34	60.03	×

$\theta_p = N/lnN$ is the TP predicted θ. $\theta_a = N/H_N$ is the actual θ.

Only Peter's actual $e < N/lnN$. Only April's $T_{Rule} < T_{NoRule}$

Table of Contents

Conclusion

On Hypothetical Data

- TP has several mathematical discrepancies that would lead to implausible results (e.g. when there are more than 97 exceptions in 100 items, a rule would be derived) or contradictory results (e.g. when there are 3 regulars in 10 items, if they rank 8,9,10, a rule can't be derived; otherwise a rule would be derived).

On Children's corpora Data

- Majority of the children's data don't conform to the TP's predictions.

Why won't TP work?

- Theoretical Assumption: a rule is derived to reduce time complexity.
  - Alternative 1: the rule is derived to reduce both time complexity and memory space.
  - Alternative 2: the rule is derived not for any utilitarian reasons.

- Operational Assumption: the time complexity has a linear relationship with the number of exceptions. The calculation relies on **dual-route model**, **serial search process**, **rank hypothesis** and **Zipfian distribution**.
  - Dual-route model vs Connectionist model
  - Serial search vs Parallel process
  - Modify t_i instead of using rank
  - Zipfian distribution doesn't really apply to small datasets

References

Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Forster, K. I. (1976). Accessing the mental lexicon. *New approaches to language mechanisms*, 257–287.

Forster, K. I. (1992). Memory-addressing mechanisms and lexical access. *Orthography, phonology, morphology, and meaning*, 413.

MacWhinney, B. (2000). *The* CHILDES P*roject: Tools for analyzing talk*. *transcription format and programs* (Vol. 1). Psychology Press.

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

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193.

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

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley Press.

Questions?

Thanks!