Testing the Tolerance Principle on Corpus Data

How well can the Tolerance Principle explain past tense overregularization?

Xiaomeng Ma, Qihui Xu, Virginia Valian, Martin Chodorow
City University of New York

INTRODUCTION

Rule learning can lead to overgeneralization errors (e.g. *He falled* [1]). But what leads to rule learning in the first place? To predict when a rule will be productive, Yang [2, 3] proposes the Tolerance Principle (TP). It captures the insight that too many exceptions make rule learning inefficient and impossible. It quantifies the theoretical threshold number of exceptions (θ) that a learner can tolerate, as in (1):

(1) $\theta = N/\ln(N)$ (2) $\theta = N/H(N)$

N is the number of types or items in the corpus that the rule is defined over and *ln* is the natural logarithm. If the exceptions *e* are no larger than θ, rule acquisition can take place.

TP has successfully predicted morphosyntactic performance in artificial language learning [4] but has not accounted for Adam's and Eve's data on past tense acquisition [3], using only verb types in children's own utterances (U_c) to represent N. Yang attributed that failure to sampling effects [3, p. 88]. Here we aim to preserve Yang's insight but develop a different version of TP which better represents the distribution of N and N itself. In particular, we replace the natural log of N with a value that provides the best fit to corpus data.

METHODOLOGY

This study evaluates and revises Yang's method and proposes a new method to compensate for sampling effects on corpus data. The new method is tested on Adam's and Eve's data.

Summary of Adam's and Eve's data

	Adam	Eve
Age of first recording	2;3	1;6
Age of first overregularized verb	2;11	1;10
No. of files in between	18	10

1. Better represent the distribution of N

In TP, estimated maximum irregularities θ are estimated using N/ln(N), assuming that the distribution of N is Zipfian. But a Zipfian distribution is not guaranteed for small sample sizes. Therefore, we measure the actual distribution of N for each child and parent and use the empirically estimated log to calculate the denominator on the right hand side of (1), namely, (2).

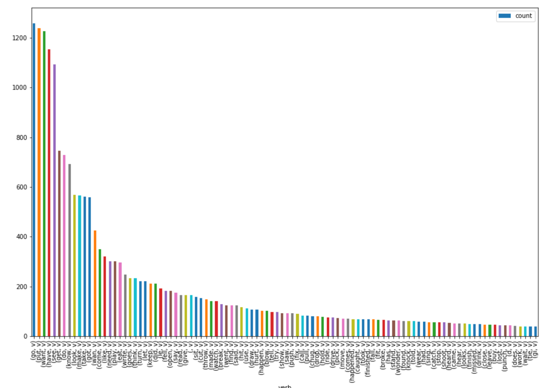

Fig.1. Adam's distribution (log=-0.64)

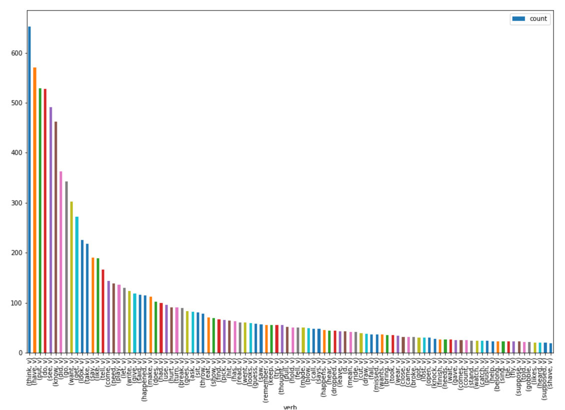

Fig. 2. Adam's mother's distribution (log = -0.72)

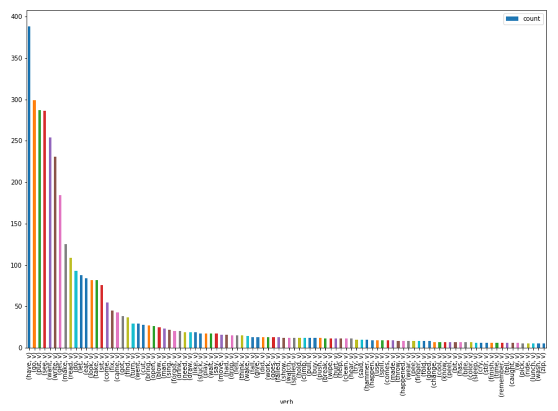

Fig. 3. Eve's distribution (log = -0.69)

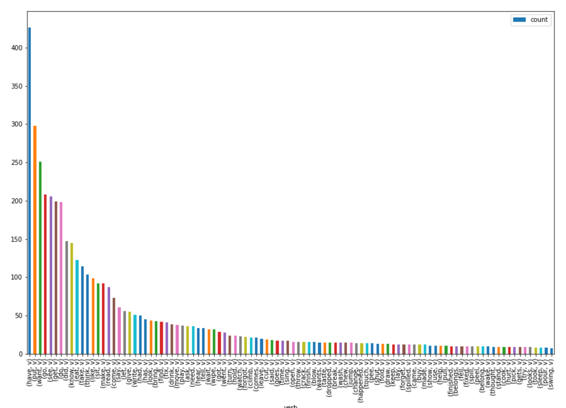

Fig. 4. Eve's mother's distribution (log = -0.68)

2. Better represent N

Instead of using children's own utterances (U_c) to represent children's effective vocabulary, we propose that *N* can be estimated through parents' input (U_p) and children's production (U_c), both of which can be extracted from 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 (%). Since λ and δ represent the comprehension cost and production loss, they should range between 0% - 100%. To estimate the minimum value of λ, we can use the proportions of words that are in the parents' utterances that are found in children's utterances. Similarly, the maximum value of δ should be equal to the maximum of λ if we make the bold assumption that children understands everything the parents said.

In addition, in order to compensate for loss of data due to undersampling, we introduce X_c and X_p for the missing data. The estimated N is shown in below:

$$(1)\ N = (U_p + X_p) \cdot \lambda = (U_c + X_c) / \delta$$

CONCLUSION

According to TP, when $\theta \geq e$, acquisition can take place. Using this method, instead of simply comparing θ and e, we evaluate the plausibility of $\theta \geq e$, namely the possible values for X_p and X_c. The number of verbs (*U*) and irregular past tense verbs (*e*) are generated automatically using the NLTK Python package.

First, we calculated the minimum value for N using the formula $\theta = N/H(N)$, by making θ=e, using the estimate of the harmonic number for parents' and children's word distributions.

Estimating N using Parent's input

	Adam	Eve
e	29	24
log	-0.72	-0.68
Estimated N	~500	~450

Estimating N using Children's output

	Adam	Eve
e	16	16
log	-0.64	-0.69
Estimated N	~320	~240

Then we insert the minimal value for N to the formula in (2) and generate an estimated value for X_p and X_c. TP is confirmed if $X_p < U_p$ and $X_c < U_c$.

Estimating X for parents' input

	Adam	Eve
U	354	229
N	~500	~450
Estimated X	146	221
X<U?	Yes	Yes

Estimating X for Children's output

	Adam	Eve
U	294	147
N	~320	~240
Estimated X	26	93
X<U?	Yes	Yes

The findings of this study show first that the distribution of verb production of is not Zipfian for either parents or children (shown in Figure 1-4) and, second, that the new method produces a plausible X. TP's predictions are confirmed for corpus data.

REFERENCE

Selected Reference:

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

[2] Yang, C. (2005). On productivity. *Linguistic variation yearbook*, *5*(1), 265-302.

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

[4] Schuler, K. D., Yang, C., & Newport, E. L. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. In *CogSci*.

contact: xma3@gradcenter.cuny.edu