Frequency Matching Behavior in Online MaxEnt Learners

USC University of Southern California

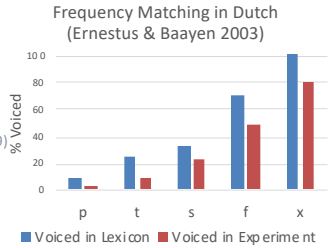
Charlie O'Hara
charleso@usc.edu

1. Overview

Not all grammatically possible patterns are equally learnable. The learnability of different patterns can be dependent on the learning algorithm.

- Human learners show *frequency matching behavior*, so our model of learning should as well.

(Ernestus & Baayen 2003, Hayes et al. 2009)



- Frequency Matching: speakers match trends observed in the lexical statistics when tested on nonce words.

Recent work has uncovered that some MaxEnt learning algorithms fail to frequency match. (Zymet 2018, 2019, Hughto et al. 2019)

CLAIM: More realistic MaxEnt can avoid these issues.
- Online learners act differently than the batch learner, never ignoring the general constraints
- With large lexicons, little weight is given to lexically specific constraints when there is no idiosyncrasy.

2. Lexical Idiosyncrasy and Frequency Matching

Lexical Idiosyncrasy: The rates at which morphemes undergo phonological processes can be morpheme-specific.

- Lexically indexed copies of every constraint, for every morpheme. (Pater 2010, Linzen et al. 2013, Hughto et al. 2019)
- Nonce words have no lexical constraints, so only general constraints apply to them.
- No unique weighting of constraints that will yield a (near-)perfect match to the observed data.

/pat/	NoCoda w=5	Max w=7	$NoCoda_{pat}$ w=5	Max_{pat} w=8	Harm	Prob
[pat]	-1		-1		-10	.993
[pa]		-1		-1	-15	.007
Nonce	**NoCoda**	**Max**	**$NoCoda_{pat}$**	**Max_{pat}**		
[VC]	-1				-5	.881
[V]		-1			-7	.119

/pat/	NoCoda w=0	Max w=0	$NoCoda_{pat}$ w=10	Max_{pat} w=15	Harm	Prob
[pat]	-1		-1		-10	.993
[pa]		-1		-1	-15	.007
Nonce	**NoCoda**	**Max**	**$NoCoda_{pat}$**	**Max_{pat}**		
[VC]	-1				0	.5
[V]		-1			0	.5

3. Abstraction in Learning Simulations

Learning simulations often abstract details of learning for purposes of computational tractability.

- Smaller representative set of lexical forms, and constraints.
- Batch learning as an approximation of online learning

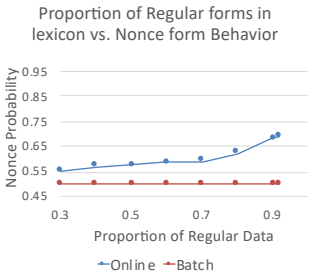
	Batch Learning	Online Learning
Method:	Learner exposed to entire dataset, maximizes probability of that data	Learner exposed to one piece of data at a time, updates weights when learner disagrees with data.
Avoids Overfitting by:	Minimize a L2 prior on constraint weights	Learner can only learn from a fixed set of errors.
Speed:	Simulations run faster	Simulations are noisier and slower
Memory	Many forms at once.	Only one form at a time

- Particularly with subtle emergent learning biases, approximations can change learning behavior!

4. Online vs. Batch Learning

Batch learners with a L2 Prior overfit the lexicon with an small simple constraint set (Zymet 2018, 2019).

- Two classes (regular and irregular) that act categorically
- Three toy constraints



regular	Undergo	Undergo (R)	Faith (I)
regular	-1	-1	
REG			
irregular	**Undergo**	**Undergo (R)**	**Faith (I)**
irregular	-1		
IRREG			-1
nonce	**Undergo**	**Undergo (R)**	**Faith (I)**
nonce	-1		
NONCE			

- Online learner shows frequency matching, batch learner does not.
- With the L2 prior, batch learners give **Undergo** a weight of zero, no matter how frequent regular forms are.
  - **Minimizes the L2 prior**
- Online learner must update **Undergo** to raise the weights of the specific constraints.

5. Lexicon Size

Hughto et al (2019) point out that batch learners with an L2 prior make **spurious use of specific constraints.**

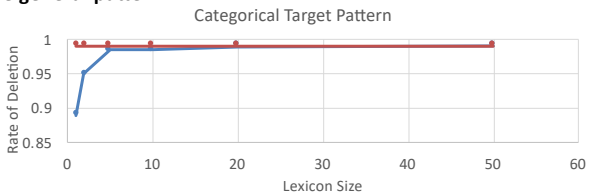
- In data without lexical idiosyncrasy, their learner makes use of lexically specific constraints, and underuses the general constraints
- As a result, the nonce-word frequencies diverge from the average aggregated across the lexicon.

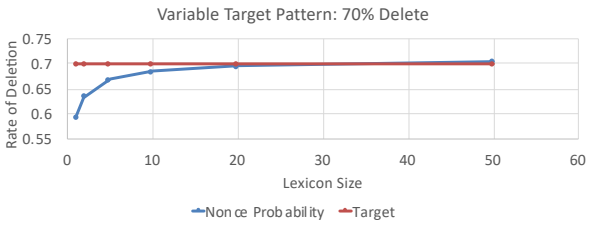
The data used in their simulations use a small toy-like lexicon

- Only three prefixes in the simulations, which each overuse their lexically specific constraints.
- Lexicon size has a major effect on how much generalization occurs.

I ran simulations using an online MaxEnt learner on a simple grammar containing NoCoda, Max, and lexically indexed versions of each constraint.

The larger the lexicon, the closer nonce-form frequency matches the general pattern.





Why? Each lexically specific constraint updates when its lexical form is sampled (1/n frequency) and an error occurs, whereas the general constraints update every time an error occurs.

Individual lexical constraints perform better than more general indexed constraints, like Undergo(Regular).

Abstractions from realistic models of learning can mask emergent biases in the learner.
Sufficient approximations of learning for some purposes can fail to match human learning in others.