Frequency Matching Behavior in On-line MaxEnt Learners - Charlie O'Hara, USC

Overview Studies have repeatedly shown that language users seem to apply processes to nonce forms at a similar rate as what is observed in the lexicon as a whole (Zuraw, 2000; Ernestus & Baayen, 2003; Hayes *et al.* , 2009; Linzen *et al.* , 2013; Moore-Cantwell, 2016; Zymet, 2018b; Hughto *et al.* , 2019) Capturing both the statistical generalizations across the whole lexicon and for individual lexical items is a challenge for MaxEnt models of phonological learning, which should be able to mimic the behavior of language users. The major challenge for learners is called THE GRAMMAR-LEXICON BALANCING PROBLEM by Zymet (2018b,a)—if the lexical constraints are too active in the grammar compared to the more general grammatical constraints, the statistical generalization across the lexicon is not captured. Most work attempting to capture such biases use batch learning algorithms, directly minimizing an objective function that balances the likelihood of capturing the training data, and a prior that limits the movement of each constraint. However, on-line error-driven learners innately exhibit a bias towards limited movement of each constraint, without any explicit prior placed on the constraint weights. I use an on-line learner to examine how the innate bias of online learners can affect the grammar-lexicon balancing problem. I find that the larger the lexicon, the closer the learner matches nonce-word frequencies to the general lexical patterns.

Background Frequency matching behaviors have been observed in experiments in a variety of languages and contexts: ranging from Tagalog nasal substitution (Zuraw, 2000), voicing alternations in Dutch (Ernestus & Baayen, 2003), to Hungarian vowel harmony (Hayes *et al.* , 2009). Several proposals have attempted to model frequency matching behaviors with MaxEnt models. Moore-Cantwell & Pater (2016) use an L2 prior on the constraint weights, and approach human behavior. Zymet (2018b) and Hughto *et al.* (2019) show that the L2 prior can make the lexical constraints too active to capture the nonce-word generalizations. Zymet (2018b) and Hughto *et al.* (2019) propose different mechanisms for solving the grammar-lexicon balancing problem, but both involve an overt prior preventing the lexical constraints from receiving too much weight. The majority of this work makes use of batch learners, however Smith & Moore-Cantwell (2017) show that an on-line learner with induced (and decaying) UR constraints performs better than batch learners at capturing allomorphy in English comparatives.

The Model The simulations here use a MaxEnt grammar with two general constraints, as well as indexed variants of both general constraints for each lexical item. These lexically indexed constraints are equivalent to the lexical scales used by Hughto *et al.* (2019), and a special case of additive scaled constraints generally (Hsu & Jesney, to appear). All constraints are limited to non-negative weights.

(1)

VC_i	MAX	NOCODA	MAX_i	$NOCODA_i$
a. VC		-1		-1
b. V	-1		-1	

I use the Perceptron learning algorithm (Rosenblatt, 1958; Boersma & Pater, 2016). On each iteration of the learning algorithm, a random lexical item is sampled from the lexicon. Output forms for that item is sampled from both the target grammar, and the learner's current grammar. These two forms are compared, if they differ, the constraints violated by the learner's incorrect output are decreased, and the constraints violated by the target grammar's output are increased. In the simulations every time a mismatch occurs between the learner and the target grammar, the two general constraints MAX and NOCODA are updated (in opposite directions); but any lexically specific constraint, say MAX_i would only be updated when an error occurred on the relevant lexical item.

Simulations To evaluate whether the learner frequency matches, I compare the rate of deletion of nonce forms to the rate of deletion averaged across all lexical forms after the learner has been exposed to a fixed amount of data.

Following Hughto *et al.* (2019), I tested several distinct types of target patterns. In all of the simulations in this paper, learners were trained on data that had at most two classes of lexical items that had the same rate of variation, presented in the table in (2). In these simulations, 60% of the items maintained their final consonants at the rates in the First Portion column, and the remaining 40% maintained their final consonants at the rates in the Second Portion column.

(2)

Pattern	First Portion	Second Portion	Target	Nonce-Rate (50 items)	Learner Average
a. Categorical	1.0		1.00	1.00	1.00
b. Variable	0.7		0.7	0.715	0.723
c. Propensity	0.7	0.3	0.54	0.509	0.559
d. Variable-Lexical	0.3	1.0	0.58	0.676	0.565
e. Lexical	1.0	0.0	0.60	0.651	0.600

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 463-465.
New Orleans, Louisiana, January 2-5, 2020

I ran simulations for each condition twenty times, starting with general markedness constraints (NoCoda) weighted at 50, and all other constraints weighted at zero, following (Tesar & Smolensky, 2000; Jesney & Tessier, 2011). Each simulation here had 50 items in the lexicon, and ran for 50,000 iterations with a learning rate of 0.1. With 50 items in the lexicon, the learner closely matches the average probability of coda consonant maintenance in the lexicon in the first three patterns, and overshoots the lexical generalization in patterns d and e, as shown in (2).

Impact of Lexicon Size To see the influence of lexicon size on the grammar-lexicon balance problem, I reran these simulations using a variety of different lexicon sizes, running each simulation for 1000 iterations per lexical item in the lexicon. Figure 1 shows that for the first three patterns, the discrepancy between the nonce-form deletion rate (solid line) and average deletion rate (dashed line) decreases monotonically as the number of lexical items increases. To see why this is, note that frequency matching occurs when the contribution of the specific constraints is minimal compared to the general constraints. The contribution of these specific constraints is dependent on how often the constraints update, and thus how often the learner observes an error on a specific lexical item. The more often the learner deletes a coda on a specific lexical item when the teacher produces it, the higher weighted Max$_i$ will be. Because learners start with markedness constraints weighted high, they will very often see deletion errors in early learning, as the general Max constraint approaches NoCoda. If the lexicon is small, the same lexical items will be chosen often and the specific constraints for those items will get too highly weighted; but if the lexicon is large, any single lexical item is unlikely to be selected too often, so most of the general phonotactic pattern is learned via updating the general constraints.

Fig 1: Frequency Matching with Larger Lexicons

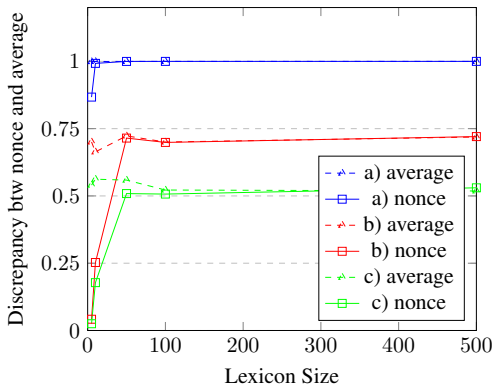

Fig 2: Overshoot with Larger Lexicons

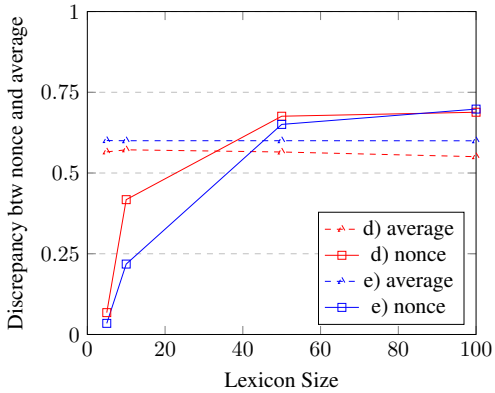

Fig 2 shows that when trained on patterns d and e, the learner overshoots the target pattern. This result resembles a bias observed by Hughto *et al.* (2019) §4.2. Further simulations showed that this overshoot is caused when one lexical class is close to, or fully categorical. In these cases, the learner learns a nonce-rate of deletion that is slightly closer to the larger categorical class's frequency than the average as a whole.

This overshoot is caused by the fact that MaxEnt grammars can never return a completely categorical mapping. The general constraints in these simulations are subject to a tug-of-war between the two lexical classes—once the general constraints have gotten close to the average mapping, the lexical idiosyncrasies must be learned. Then, if a lexical item from the class with a lower rate of deletion is sampled, Max and the relevant specific Max$_i$ are increased (and the NoCoda constraints are decreased). Then, when a lexical item from the class with a greater rate of deletion is sampled, the general constraints shift back, and the relevant specific NoCoda$_j$ is increased. If one class is smaller, it's rate of deletion will be further from the average, so each time one of its forms are sampled, it will be more likely to cause an error. Most updates on a specific lexical item will be in one direction, but if that item is variable it is possible that the learner observes updates the opposite direction, either by chance, or because the learner overshot the correct rate of deletion for that item. These updates in the opposite direction help ensure that as more of the lexical forms are learned, the amount they pull on the general constraints decreases. However, if one lexical class is categorical, the learner can never overshoot the target rate of deletion for that class. There will always be a minute tug from every lexical item in the categorical class on the general constraints. With a larger lexicon, these minute tugs compound, leading to the type of overshoot seen in Figure 2.

Without any overt priors to keep the lexical constraints from capturing much weight, on-line MaxEnt learners exhibit frequency matching behavior in most conditions.

References

Boersma, Paul, & Pater, Joe. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. *In:* McCarthy, John J., & Pater, Joe (eds), *Harmonic Grammar and Harmonic Serialism*. Equinox.

Ernestus, Mirjam, & Baayen, R. Harald. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 5–38.

Hayes, Bruce, Zuraw, Kie, Siptar, Peter, & Londe, Zsuzsa. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, **85**, 822–863.

Hsu, Brian, & Jesney, Karen. to appear. Scalar Positional Markedness and Faithfulness in Harmonic Grammar. *In: Proceedings of CLS 51*.

HUGHTO, CORAL, LAMONT, ANDREW, PRICKETT, BRANDON, & JAROSZ, GAJA. 2019. Learning Exceptionality and Variation with Lexically Scaled MaxEnt. *In: Proceedings of the Society for Computation in Linguistics*, vol. 2.
JESNEY, KAREN, & TESSIER, ANNE-MICHELLE. 2011. Biases in Harmonic Grammar: The road to restrictive learning. *Natural Language & Linguistic Theory*, **29**.
LINZEN, TAL, KASYANENKO, SOFYA, & GOUSKOVA, MARIA. 2013. Lexical and phonological variation in Russian preopositions. *Phonology*, **30**, 453–515.
MOORE-CANTWELL, CLAIRE. 2016. *The representation of probabilistic phonological patterns: Neurological, behavioral, and computational evidence from the English stress system*. Ph.D. thesis, University of Massachusetts Amherst.
MOORE-CANTWELL, CLAIRE, & PATER, JOE. 2016. Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints. *Catalan Journal of Linguistics*, **15**.
ROSENBLATT, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408.
SMITH, BRIAN, & MOORE-CANTWELL, CLAIRE. 2017. Emergent idiosyncrasy in English comparatives. *In:* LAMONT, ANDREW, & TETZLOFF, KATIE (eds), *NELS 47: Proceedings of the 47th meeting of the North East Linguistic Society*.
TESAR, BRUCE, & SMOLENSKY, PAUL. 2000. *Learnability in Optimality Theory*. MIT Press.
ZURAW, KIE. 2000. *Patterned Exceptions in Phonology*. Ph.D. thesis, UCLA, Los Angeles.
ZYMET, JESSE. 2018a. Learning a Frequency-Matching Grammar together with Lexical Idiosyncrasy: MaxEnt versus Hierarchical Regression. *In:* HOUT, KATHERINE, MAI, ANNA, MCCOLLUM, ADAM, ROSE, SHARON, & ZASLANSKY, MATT (eds), *Proceedings of the 2018 Annual Meeting on Phonology*. Washington, DC: Linguistics Society of America.
ZYMET, JESSE. 2018b. *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning*. Ph.D. thesis, University of California Los Angeles.