Emergent avoidance of cumulativity and variability in phonological typology

Coral Hughto

University of Massachusetts Amherst

Workshop on Analyzing Typological Structure
Stanford - 22 September 2018

Introduction

- Traditional goal of typology: predict divide between attested and unattested patterns
  - Grammatical model should be able to represent all and only attested patterns
- Potential issue: traditional focus on *categorical* grammar models and patterns
- What do I mean by "probabilistic typology"?
  - Frequency predictions over patterns in a typology
    - e.g. Pater 2012, Staubs 2014, Stanton 2016, O'Hara 2018, this work, among others

In this talk

- I examine the effects of combining a grammatical model with agent-based learning models on frequency predictions over patterns in a typology
- Method draws on differences in learnability to explain differences in frequency
- Grammatical model: Maximum Entropy (MaxEnt; Goldwater & Johnson 2003)
- Learning models: Interactive learning and Iterated learning
- Learning biases observed in both learning models:
  - Bias away from constraint cumulativity (gang effects)
  - Bias away from variability (such that agents accumulate majority probability on one output per input)

MaxEnt

- Uses weighted constraints; defines a probability distribution over competing output candidates

/In_1/	3 X	2 Y	H	p
→A		-1	-2	0.73
B	-1		-3	0.27

/In_2/	3 X	2 Y	H	p
→C	-1		-3	0.73
D		-2	-4	0.27

- Harmony score (H) = weighted sum of constraint violations
  - $H(x) = \sum_{i=1}^{n} W(C_i) * C_i(x)$
- Probability (p) = proportion of exponentiated Harmony out of sum over competing candidate set
  - $p(x) = \frac{e^{H(x)}}{e^{H(x)} + e^{H(y)} + e^{H(z)} \ldots}$

Gang Effects

/*In*$_1$/	3 X	2 Y	*H*	*p*
→A		-1	-2	0.73
B	-1		-3	0.27

/*In*$_2$/	3 X	2 Y	*H*	*p*
→C	-1		-3	0.73
D		-2	-4	0.27

- Weighted constraint grammars allow for cumulative constraint interaction (a.k.a. gang effects)
- Multiple violations of (a) lower-weighted constraint(s) can cumulatively outweigh one violation of a higher-weighted constraint

Do we need gang effects?

/*In*$_1$/	3 X	2 Y	*H*	*p*
→A		-1	-2	0.73
B	-1		-3	0.27

/*In*$_2$/	3 X	2 Y	*H*	*p*
→C	-1		-3	0.73
D		-2	-4	0.27

- This property of weighted constraint grammars has been criticized for overpredicting the space of typological possibilities (e.g. Legendre et al. 2006, but see Pater 2009)
- However, gang effects can be useful in representing:
  - stress windows (Staubs 2014)
  - "general-case" neutralization (Hughto & Pater 2017)

The effect of learning?

- *Assuming* some patterns are represented by gang effects (see e.g. Hughto & Pater 2017, Zuraw & Hayes 2017):
  - What kind of predictions does MaxEnt make about the predicted frequency of gang effect patterns?
  - What effect might learning have on the predicted probabilistic typology?
- Paired MaxEnt with each of two agent-based learning models, investigating different mechanisms that may influence acquisition
- Interactive learning model: simulates the effect of mutual peer interaction/influence on language development
- Iterated learning model: simulates the effect of transmitting a language across generations

Interactive Learning - $A_1 \leftrightarrow A_2$

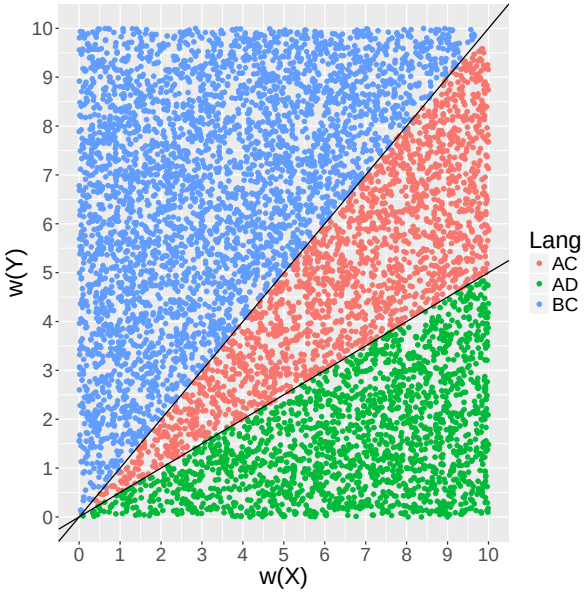
- Two agents take turns in the roles of teacher and learner
- In each run, the agents exchange data for some number of learning steps
- Error-driven Perceptron learning algorithm
  - Teacher samples input, output based on its current grammar
  - Learner samples an output for same input
  - If the agents don't agree, the learner updates its grammar
  - New Weights=OId Weights+(Teacher-Learner)*Learning Rate
- The distribution over patterns and probabilities across multiple runs is taken as the predicted typology

Iterated Learning - $A_1 \rightarrow A_2$, $A_2 \rightarrow A_3$, $A_3 \rightarrow A_4$ ...

- Approximates the transmission of a language across generations
- One agent serves as the "teacher" (the target grammar) for a "learner" agent
- After some number of learning steps, the learner becomes the teacher for a new learner
  - Same error-driven learning algorithm
- Process repeats for some number of generations
- The distribution over patterns and probabilities across multiple runs is taken as the predicted typology

Minimal Working Example

/In_1/	X	Y
A		-1
B	-1	
/In_2/		
D		-2
C	-1	



- Three possible patterns:
  - BC : w(Y) > w(X)
  - AD : w(X) > 2w(Y)
  - AC : 2w(Y) > w(X) > w(Y) (Gang effect)
- Pattern = the set of highest probability output candidates

Minimal Working Example

/In_1/	1 X	3 Y	H
A		-1	-3
→B	-1		-1
/In_2			
→C	-1		-1
D		-2	-6

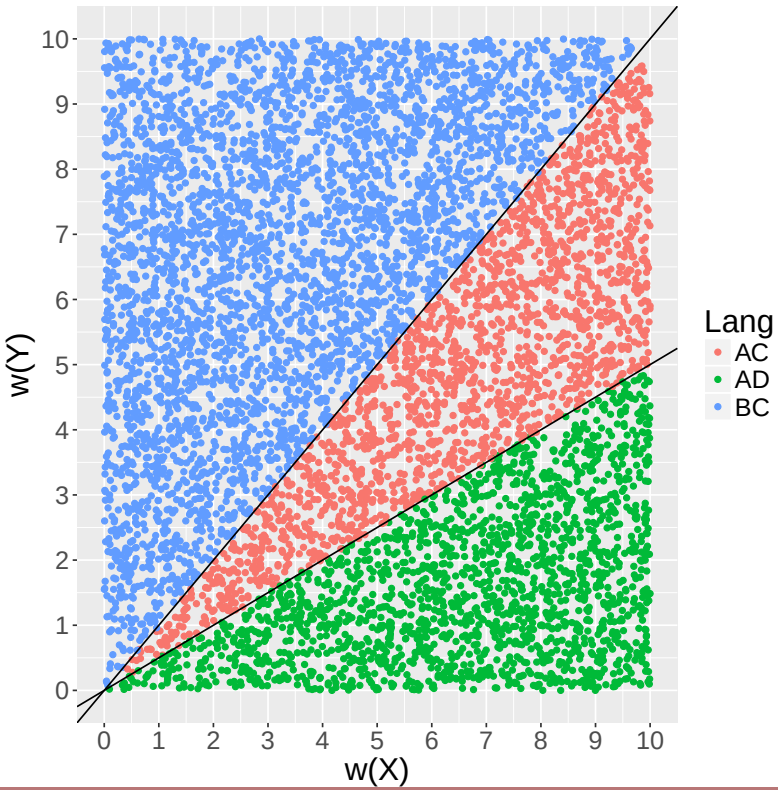
/In_1/	3 X	1 Y	H
→A		-1	-1
B	-1		-3
/In_2			
C	-1		-3
→D		-2	-2

/In_1/	3 X	2 Y	H
→A		-1	-2
B	-1		-3
/In_2			
→C	-1		-3
D		-2	-4

- Three possible patterns:
  - BC : w(Y) > w(X)
  - AD : w(X) > 2w(Y)
  - AC : 2w(Y) > w(X) > w(Y) (Gang effect)

Results - Before Learning

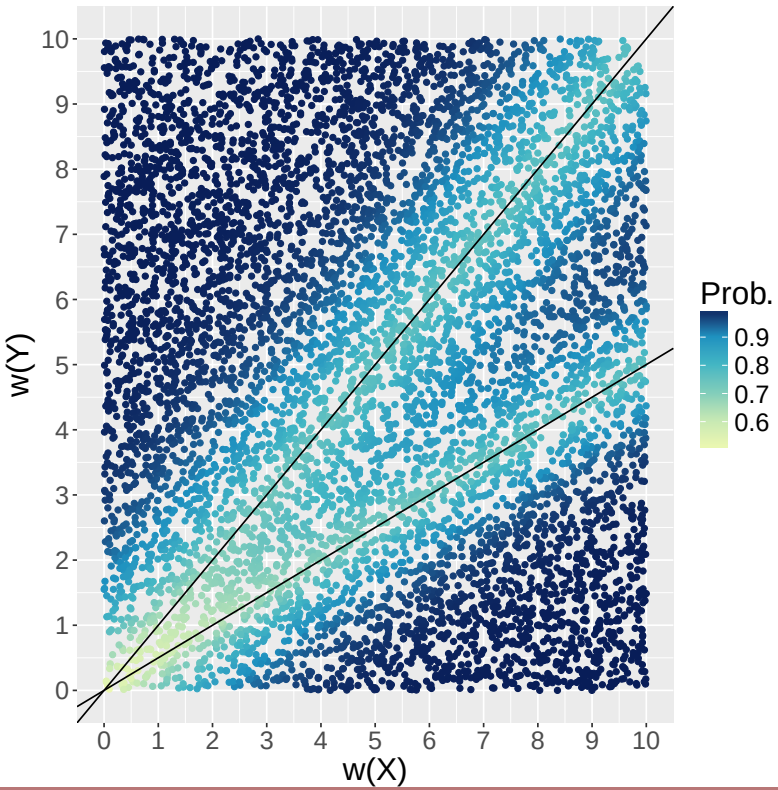
- Baseline distribution obtained by sampling 1,500 pairs of constraint weights from a uniform distribution (0.0-10.0)
- Similar in concept to r-volume (Riggle 2010)



Pattern	Proportion
BC	0.50
AD	0.25
AC (gang)	0.25
Avg. Prob.	**Proportion**
0.9-1.0	0.60
0.8-0.9	0.22
0.7-0.8	0.15
0.6-0.7	0.02
0.5-0.6	0.01

Results

- Baseline distribution obtained by sampling 1,000 pairs of constraint weights from a uniform distribution (0.0-10.0)
- Similar in concept to r-volume (Riggle 2010)



Pattern	Proportion
BC	0.50
AD	0.25
AC	0.25
Avg. Prob.	**Proportion**
0.9-1.0	0.60
0.8-0.9	0.22
0.7-0.8	0.15
0.6-0.7	0.02
0.5-0.6	0.01

Model Parameters

- Interactive Learning Model:
  - 1500 runs, 5000 learning steps
- Iterated Learning Model:
  - 1500 runs, 1000 learning steps, 50 generations
- Learning rate: 0.1
- Conditions for initial weights
  - Zero
  - Randomly sampled from uniform distribution between 0.0-10.0
- Interactive model (Random): Initial states balanced across runs for pattern type and probability bin
- Iterated model (both): Initial target states randomly sampled and balanced across runs
- Iterated model (Random): unbalanced

Results

- Both models decrease the predicted frequency of the gang effect AC pattern, relative to the baseline
- Effect is more extreme for Zero than Random
- Effects are least robust for Random-Iterated

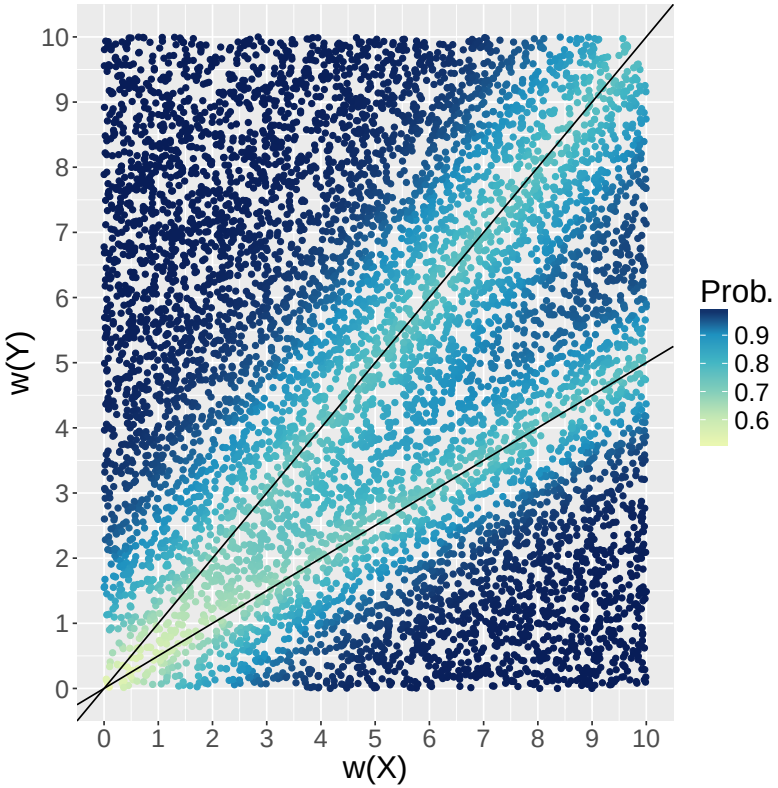
		Zero		Random	
Pattern	Baseline	Interactive	Iterated	Interactive	Iterated
BC	0.50	0.70	0.57	0.55	0.43
AD	0.25	0.25	0.40	0.34	0.36
AC	0.25	0.04	0.03	0.11	0.21

Results

- Both models also show a bias towards more categorical patterns

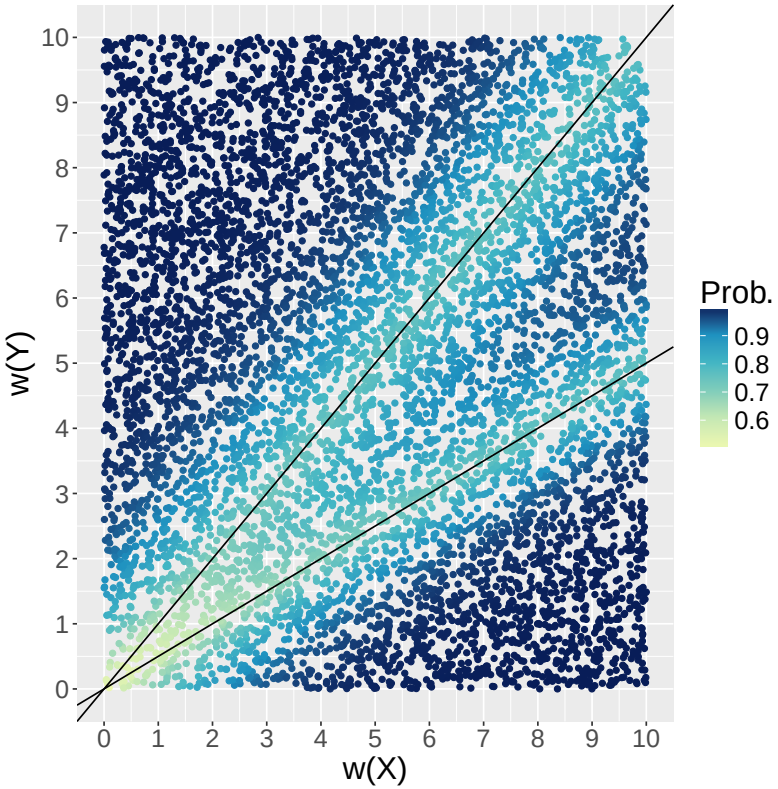
		Zero		Random	
Pattern	Baseline	Interactive	Iterated	Interactive	Iterated
0.9-1.0	0.60	0.81	0.74	0.84	0.71
0.8-0.9	0.22	0.09	0.14	0.10	0.20
0.7-0.8	0.15	0.07	0.06	0.06	0.10
0.6-0.7	0.02	0.03	0.05	0.01	0.01
0.5-0.6	0.01	0.01	0.01	0.01	0.00

Discussion - Interactive Learning Model



- Biases: away from gang effect (AC) pattern; away from variation
- Variation: Agents less likely to move out of grammars with more peaked probability distributions
- Gang effect (AC) region contains more variation than other regions

Discussion - Iterated Learning Model



- Biases: away from gang effect (AC) pattern, more extreme with initial weights of zero; away from variation
- Agents more likely to mislearn into a grammar with more peaked probability distribution
- Gang effect (AC) region contains more variation
- Agents initialized with random weights often need to pass through central space to reach their target

Discussion

- I showed how combining (a) grammatical model(s) with (a) learning model(s) can influence the predicted frequencies of patterns in the typology
- The Interactive and Iterated learning models both produced biases:
  - Away from the gang effect (AC) pattern
  - Towards more categorical patterns
- Here, I looked at the effects on pattern type (focusing on gang effects) and variability, but not (yet) at the interaction

Thanks!

Thanks to Joe Pater, Gaja Jarosz, UMass Sound Workshop, audiences at MfM, AMP, SCiL, and other workshops and conferences, and everyone here!

NOTES TO SELF

- Shift focus away from "let's get rid of gang effects!" more to my usual "how do these learning models affect the probabilistic typological predictions?"
- It's not clear in the current iteration of my presentation why I'm looking at two/both learning models; give a little detail about the different learning mechanisms and parallels/relation to real-world mechanics
- Be more explicit, and more explicit earlier, that I'm assuming that we want gang effects, and want them to be rare (cite myself!)