Generating gradient typological predictions with an interactive learning model

Coral Hughto

University of Massachusetts Amherst

PhoNE at UMass
8 April 2017

Background

- Traditionally, work in phonological typology has focused on predicting the division between possible patterns and impossible patterns

Background

- Traditionally, work in phonological typology has focused on predicting the division between possible patterns and impossible patterns
  - E.g. OT factorial typology

Background

- Traditionally, work in phonological typology has focused on predicting the division between possible patterns and impossible patterns
  - E.g. OT factorial typology
- This approach, by design, abstracts over observed differences in attested frequency between different patterns

Background

- Traditionally, work in phonological typology has focused on predicting the division between possible patterns and impossible patterns
  - E.g. OT factorial typology
- This approach, by design, abstracts over observed differences in attested frequency between different patterns
  - Grammar is about possible representations; something extra is needed to motivate frequency differences

Background

- Traditionally, work in phonological typology has focused on predicting the division between possible patterns and impossible patterns
  - E.g. OT factorial typology
- This approach, by design, abstracts over observed differences in attested frequency between different patterns
  - Grammar is about possible representations; something extra is needed to motivate frequency differences
- We can generate predictions about the frequency of possible patterns by combining a grammar theory with a learning model

Background

- Traditionally, work in phonological typology has focused on predicting the division between possible patterns and impossible patterns
  - E.g. OT factorial typology
- This approach, by design, abstracts over observed differences in attested frequency between different patterns
  - Grammar is about possible representations; something extra is needed to motivate frequency differences
- We can generate predictions about the frequency of possible patterns by combining a grammar theory with a learning model
- This work combines Maximum Entropy (MaxEnt) grammar with an agent-based, interactive learning model

Gang effects

- MaxEnt uses weighted constraints, and defines a probability distribution over output candidates for each input

Gang effects

- MaxEnt uses weighted constraints, and defines a probability distribution over output candidates for each input
- Weighted constraints allow for "gang effects": patterns where multiple violations of lower-weighted constraints outweigh one violation of a higher weighted constraint

Gang effects

- MaxEnt uses weighted constraints, and defines a probability distribution over output candidates for each input
- Weighted constraints allow for "gang effects": patterns where multiple violations of lower-weighted constraints outweigh one violation of a higher weighted constraint

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

Gang effects

- We may want this extra representational power, even though it can predict unattested patterns, because we can model some attested patterns as gang effects

Gang effects

- We may want this extra representational power, even though it can predict unattested patterns, because we can model some attested patterns as gang effects
- "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere

/sa/	→	[sa]		/ʃa/	→	[sa]
/si/	→	[si]		/ʃi/	→	[ʃi]

General-Case Neutralization (GCN) grammar

weights	3	2	2	
/sa/	NO[ʃ]	NO[si]	IDENT	
☞sa				0
ʃa	-1		-1	-5
/ʃa/	NO[ʃ]	NO[si]	IDENT	
☞sa			-1	-2
ʃa	-1			-3
/si/	NO[ʃ]	NO[si]	IDENT	
☞si		-1		-2
ʃi	-1		-1	-5
/ʃi/	NO[ʃ]	NO[si]	IDENT	
si		-1	-1	-4
☞ʃi	-1			-3

Recap

- We want to:

Recap

- We want to:
  - Generate gradient typological predictions to account for frequency differences between attested phonological patterns

Recap

- We want to:
  - Generate gradient typological predictions to account for frequency differences between attested phonological patterns
  - Constrain the overprediction of gang effects in weighted constraint grammars, while allowing their possibility

Recap

- We want to:
  - Generate gradient typological predictions to account for frequency differences between attested phonological patterns
  - Constrain the overprediction of gang effects in weighted constraint grammars, while allowing their possibility
- Solution to both: an agent-based, interactive learning model
  - Paired with a MaxEnt grammar

Recap

- We want to:
  - Generate gradient typological predictions to account for frequency differences between attested phonological patterns
  - Constrain the overprediction of gang effects in weighted constraint grammars, while allowing their possibility
- Solution to both: an agent-based, interactive learning model
  - Paired with a MaxEnt grammar
- This learning model generates gradient typological predictions, while reducing the predicted probability of gang effect patterns (relative to simply sampling from the space of possible grammars)

The Model

- Interactive learning model: two agents exchange data and between themselves generate a language (e.g. Dediu 2009, Pater & Moreton 2012)

The Model

- Interactive learning model: two agents exchange data and between themselves generate a language (e.g. Dediu 2009, Pater & Moreton 2012)
  - $Agent_1 \leftrightarrow Agent_2$

The Model

- Interactive learning model: two agents exchange data and between themselves generate a language (e.g. Dediu 2009, Pater & Moreton 2012)
  - $Agent_1 \leftrightarrow Agent_2$
  - No target language; agents take turns being "teacher" and "learner"

The Model

- Interactive learning model: two agents exchange data and between themselves generate a language (e.g. Dediu 2009, Pater & Moreton 2012)
  - $Agent_1 \leftrightarrow Agent_2$
  - No target language; agents take turns being "teacher" and "learner"
- Contrast with an iterated learning model (e.g. Kirby & Hurford 2002) where an agent learns from a target distribution, then becomes teacher to the next agent in the learning chain

The Model

- Interactive learning model: two agents exchange data and between themselves generate a language (e.g. Dediu 2009, Pater & Moreton 2012)
  - $Agent_1 \leftrightarrow Agent_2$
  - No target language; agents take turns being "teacher" and "learner"
- Contrast with an iterated learning model (e.g. Kirby & Hurford 2002) where an agent learns from a target distribution, then becomes teacher to the next agent in the learning chain
  - $Agent_1 \rightarrow Agent_2$

The Model

- Interactive learning model: two agents exchange data and between themselves generate a language (e.g. Dediu 2009, Pater & Moreton 2012)
  - $Agent_1 \leftrightarrow Agent_2$
  - No target language; agents take turns being "teacher" and "learner"
- Contrast with an iterated learning model (e.g. Kirby & Hurford 2002) where an agent learns from a target distribution, then becomes teacher to the next agent in the learning chain
  - $Agent_1 \rightarrow Agent_2$
  - $Agent_2 \rightarrow Agent_3$

How it works

- Two agents begin in an initial state (e.g. zero weights or random weights) with a given set of tableaux

How it works

- Two agents begin in an initial state (e.g. zero weights or random weights) with a given set of tableaux
- Agents interact (exchange data) for a number of learning steps

How it works

- Two agents begin in an initial state (e.g. zero weights or random weights) with a given set of tableaux
- Agents interact (exchange data) for a number of learning steps
- From initial state to final learning step = 1 run of the simulation

How it works

- Two agents begin in an initial state (e.g. zero weights or random weights) with a given set of tableaux
- Agents interact (exchange data) for a number of learning steps
- From initial state to final learning step = 1 run of the simulation
- The distribution of languages learned over a given number of runs is taken as the predicted probability distribution over languages

How it works

- Two agents begin in an initial state (e.g. zero weights or random weights) with a given set of tableaux
- Agents interact (exchange data) for a number of learning steps
- From initial state to final learning step = 1 run of the simulation
- The distribution of languages learned over a given number of runs is taken as the predicted probability distribution over languages
- Languages that the agents learn more often are predicted to occur more often typologically

How it works

- In each learning step:

How it works

- In each learning step:
  - One agent becomes "teacher", the other is "learner"

How it works

- In each learning step:
  - One agent becomes "teacher", the other is "learner"
  - An input is randomly selected, and each agent samples an output according to its current grammar

How it works

- In each learning step:
  - One agent becomes "teacher", the other is "learner"
  - An input is randomly selected, and each agent samples an output according to its current grammar
  - The learner compares its output to the teacher's

How it works

- In each learning step:
  - One agent becomes "teacher", the other is "learner"
  - An input is randomly selected, and each agent samples an output according to its current grammar
  - The learner compares its output to the teacher's
  - If they are different, the learner updates its constraint weights

How it works

- In each learning step:
  - One agent becomes "teacher", the other is "learner"
  - An input is randomly selected, and each agent samples an output according to its current grammar
  - The learner compares its output to the teacher's
  - If they are different, the learner updates its constraint weights
  - Roles reverse

How it works

- Constraint weight update:

How it works

- Constraint weight update:
- New Weights = Old Weights +
  (Teacher's Violations - Learner's Violations) * Learning Rate

How it works

- Constraint weight update:
- New Weights = Old Weights +
  (Teacher's Violations - Learner's Violations) * Learning Rate
- The update promotes constraints that favor the teacher's output, and demotes constraints favoring the learner's output

How it works

- Constraint weight update:
- New Weights = Old Weights +
  (Teacher's Violations - Learner's Violations) * Learning Rate
- The update promotes constraints that favor the teacher's output, and demotes constraints favoring the learner's output
- The agents can't see each others' grammars, only the outputs they produce

Minimal Working Example

- As a tiny example to illustrate, consider the tableaux below:

	X	Y
A		-1
B	-1	
D		-2
C	-1	

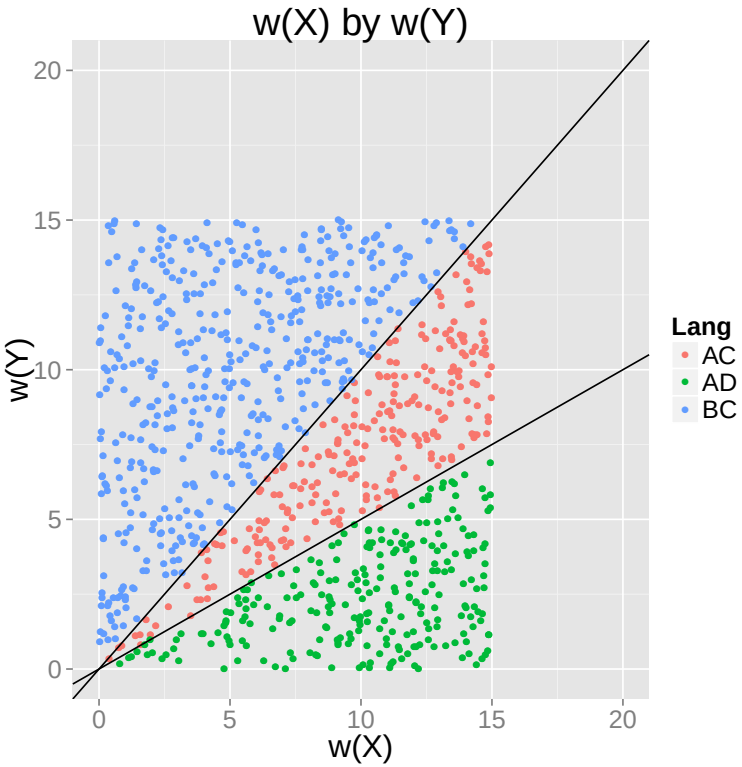
- Three possible languages:
  - BC : $w(Y) > w(X)$
  - AD : $w(X) > 2w(Y)$
  - AC : $2w(Y) > w(X) > w(Y)$ (Gang effect!)

Learning Simulations

- 1,000 runs of the Interactive Learning Model were performed
- Initial agent weights were randomly sampled from a uniform distribution with the range 0-15
- Agents interacted for 30,000 learning steps
- Learning rate was 0.1

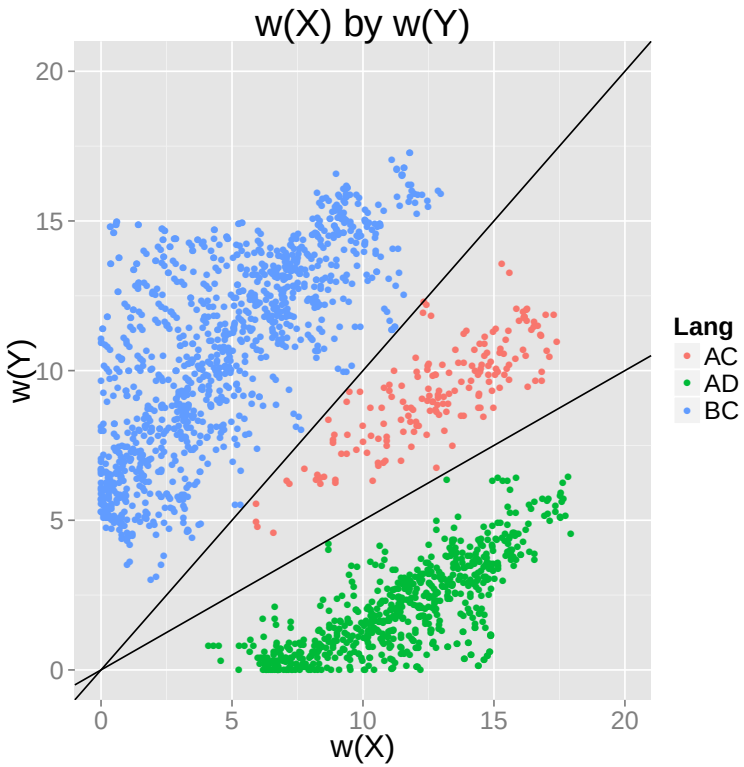
Distribution over initial weights, before interaction

- We can take the initial distribution as a baseline prediction
- BC: 49% ; AD: 25% ; AC: 26%



Predicted distribution, after interaction

- After 30,000 learning steps, predicted probability of gang effect AC language has decreased
- BC: 57% ; AD: 35% ; AC: 9%



Effects of the Interactive Learning Model

- In the Interactive Learning Model, agents interact, exchanging data by sampling outputs from their current grammar

Effects of the Interactive Learning Model

- In the Interactive Learning Model, agents interact, exchanging data by sampling outputs from their current grammar
- The agents' constraint weights tend to move away from the borders between the possible languages

Effects of the Interactive Learning Model

- In the Interactive Learning Model, agents interact, exchanging data by sampling outputs from their current grammar
- The agents' constraint weights tend to move away from the borders between the possible languages
- This has the effect of:

Effects of the Interactive Learning Model

- In the Interactive Learning Model, agents interact, exchanging data by sampling outputs from their current grammar
- The agents' constraint weights tend to move away from the borders between the possible languages
- This has the effect of:
  - Avoiding variation in the grammar (probability accumulates on one output candidate per input)

Effects of the Interactive Learning Model

- In the Interactive Learning Model, agents interact, exchanging data by sampling outputs from their current grammar
- The agents' constraint weights tend to move away from the borders between the possible languages
- This has the effect of:
  - Avoiding variation in the grammar (probability accumulates on one output candidate per input)
  - Reducing the predicted probability of gang effects

Test Case: Palatalization

- Testing predictions: palatalization typology

Test Case: Palatalization

- Testing predictions: palatalization typology
- Earlier: "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere

Test Case: Palatalization

- Testing predictions: palatalization typology
- Earlier: "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere
- Interaction between constraints: No[ʃ], No[si], Ident

Test Case: Palatalization

- Testing predictions: palatalization typology
- Earlier: "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere
- Interaction between constraints: No[ʃ], No[si], Ident
- With these constraints, 5 possible languages:

Test Case: Palatalization

- Testing predictions: palatalization typology
- Earlier: "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere
- Interaction between constraints: NO[ʃ], NO[si], IDENT
- With these constraints, 5 possible languages:
  - (44%) Total Neutralization (TN)

Test Case: Palatalization

- Testing predictions: palatalization typology
- Earlier: "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere
- Interaction between constraints: NO[ʃ], NO[si], IDENT
- With these constraints, 5 possible languages:
  - (44%) Total Neutralization (TN)
  - (37%) Full Contrast (FC)

Test Case: Palatalization

- Testing predictions: palatalization typology
- Earlier: "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere
- Interaction between constraints: NO[ʃ], NO[si], IDENT
- With these constraints, 5 possible languages:
  - (44%) Total Neutralization (TN)
  - (37%) Full Contrast (FC)
  - (10.3%) Complementary Distribution (CD)

Test Case: Palatalization

- Testing predictions: palatalization typology
- Earlier: "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere
- Interaction between constraints: NO[ʃ], NO[si], IDENT
- With these constraints, 5 possible languages:
  - (44%) Total Neutralization (TN)
  - (37%) Full Contrast (FC)
  - (10.3%) Complementary Distribution (CD)
  - (8.2%) Special-Case Neutralization (SCN) (contrast neutralized before e.g. high vowels)

Test Case: Palatalization

- Testing predictions: palatalization typology
- Earlier: "General-case Neutralization" in Gujarati (Carroll 2012 (ms.)): /s/ and /ʃ/ contrast before /i/, but neutralize to /s/ elsewhere
- Interaction between constraints: No[ʃ], No[si], Ident
- With these constraints, 5 possible languages:
  - (44%) Total Neutralization (TN)
  - (37%) Full Contrast (FC)
  - (10.3%) Complementary Distribution (CD)
  - (8.2%) Special-Case Neutralization (SCN) (contrast neutralized before e.g. high vowels)
  - (0.5%) General-Case Neutralization (GCN) (contrast maintained before e.g. high vowels but neutralized elsewhere)

Results

- Zero: Agents initialized with constraint weights at zero
- Random: Agents initialized with sampled weights, 0-10
- Sampling: Just sampling constraint weights, no interaction

Type	Observed	Zero	Random	Sampling
Total Neut.	44%	46.6%	25.7%	16.8%
Full Contrast	37%	48%	47.5%	41.3%
Comp. Dist.	10.3%	2.6%	7.7%	8.3%
Contextual Neut.	8.2%	2.7%	8%	8.4%
General-case Neut.	0.5%	0.1%	11.1%	25%
r^2		0.96	0.63	0.17

Interim discussion

- With the palatalization case, we saw:

Interim discussion

- With the palatalization case, we saw:
  - Reduction of the predicted probability of the gang effect pattern (more so in Zero condition than Random)

Interim discussion

- With the palatalization case, we saw:
  - Reduction of the predicted probability of the gang effect pattern (more so in Zero condition than Random)
  - Successful modeling of the frequency difference between Total Neut./Full Contrast and Comp. Dist./Contextual Neut.

Interim discussion

- With the palatalization case, we saw:
  - Reduction of the predicted probability of the gang effect pattern (more so in Zero condition than Random)
  - Successful modeling of the frequency difference between Total Neut./Full Contrast and Comp. Dist./Contextual Neut.
- Difficulty: estimating observed pattern frequencies

Interim discussion

- With the palatalization case, we saw:
  - Reduction of the predicted probability of the gang effect pattern (more so in Zero condition than Random)
  - Successful modeling of the frequency difference between Total Neut./Full Contrast and Comp. Dist./Contextual Neut.
- Difficulty: estimating observed pattern frequencies
- Easier test case?: Syllable structure

Syllable Structure Data

- This effort is still preliminary: I have no numerical estimates

Syllable Structure Data

- This effort is still preliminary: I have no numerical estimates
- According to the WALS chapter on syllable structure:

Syllable Structure Data

- This effort is still preliminary: I have no numerical estimates
- According to the WALS chapter on syllable structure:
  - Languages with "moderately complex" syllable structure (up to CCVC) are more common than languages with "simple" syllable structure (up to CV)

Syllable Structure Data

- This effort is still preliminary: I have no numerical estimates
- According to the WALS chapter on syllable structure:
  - Languages with "moderately complex" syllable structure (up to CCVC) are more common than languages with "simple" syllable structure (up to CV)
  - (C)V languages are more common than CV languages

Syllable Structure Model: Set-up

- Following Bane & Riggle (2009), I use five constraints:

Syllable Structure Model: Set-up

- Following Bane & Riggle (2009), I use five constraints:
  - Onset, NoCoda, Max, Dep-V, Dep-C

Syllable Structure Model: Set-up

- Following Bane & Riggle (2009), I use five constraints:
  - Onset, NoCoda, Max, Dep-V, Dep-C
- No complex onsets or codas

Syllable Structure Model: Set-up

- Following Bane & Riggle (2009), I use five constraints:
  - Onset, NoCoda, Max, Dep-V, Dep-C
- No complex onsets or codas
- All syllables must contain a V (nucleus)

Syllable Structure Model: Set-up

- Following Bane & Riggle (2009), I use five constraints:
  - Onset, NoCoda, Max, Dep-V, Dep-C
- No complex onsets or codas
- All syllables must contain a V (nucleus)
- With weighted constraints, 5 possible languages (abstracting over repair strategies):

Syllable Structure Model: Set-up

- Following Bane & Riggle (2009), I use five constraints:
  - Onset, NoCoda, Max, Dep-V, Dep-C
- No complex onsets or codas
- All syllables must contain a V (nucleus)
- With weighted constraints, 5 possible languages (abstracting over repair strategies):

Allowed Syllables	Description	
CV	Onsets required, codas banned	
(C)V	Onsets optional, codas banned	
CV(C)	Onsets required, codas allowed	
(C)V(C)	Onsets optional, codas allowed	
CV(C) & VC	No [V] syllables	(Gang effect)

Model Results

- Observed data from WALS:
  - "moderately complex" (up to CCVC) > "simple" (CV or (C)V)
  - (C)V > CV

Model Results

- Observed data from WALS:
  - "moderately complex" (up to CCVC) > "simple" (CV or (C)V)
  - (C)V > CV
- Model predictions:
  - "simple" > "moderately complex"
  - CV > (C)V

Language	Predicted Prob.
CV	0.43
(C)V	0.25
CV(C)	0.16
(C)V(C)	0.15
CV(C);VC	0.01

Discussion

- Following the goals outlined in the introduction, the Interactive Learning Model:

Discussion

- Following the goals outlined in the introduction, the Interactive Learning Model:
  - Generates gradient typological predictions

Discussion

- Following the goals outlined in the introduction, the Interactive Learning Model:
  - Generates gradient typological predictions
  - Reduces the predicted probability of patterns which rely on gang effect interactions

Discussion

- Following the goals outlined in the introduction, the Interactive Learning Model:
  - Generates gradient typological predictions
  - Reduces the predicted probability of patterns which rely on gang effect interactions
- Promising results for palatalization typology

Discussion

- Following the goals outlined in the introduction, the Interactive Learning Model:
  - Generates gradient typological predictions
  - Reduces the predicted probability of patterns which rely on gang effect interactions
- Promising results for palatalization typology
- However, for the syllable structure test case, the model's typological predictions do not fit the descriptive observations given in WALS

Discussion

- Following the goals outlined in the introduction, the Interactive Learning Model:
  - Generates gradient typological predictions
  - Reduces the predicted probability of patterns which rely on gang effect interactions
- Promising results for palatalization typology
- However, for the syllable structure test case, the model's typological predictions do not fit the descriptive observations given in WALS
- It seems that in the case of syllable structure, there is some pressure that is not being captured here

Further steps

- Gather more detailed observed typological frequency data

Further steps

- Gather more detailed observed typological frequency data
- Investigate other biases that might be added to this system?

Further steps

- Gather more detailed observed typological frequency data
- Investigate other biases that might be added to this system?
- Other good test cases?