Emergence of strict domination effects with weighted constraints

Coral Hughto and Joe Pater

University of Massachusetts Amherst

CLS Dynamic Modeling Workshop
24 May 2017

In This Talk

- Introduce a model of typology which incorporates learning through agent-based modeling
  - Maximum Entropy grammar with an agent-based, interactive learning model
- Show how the effects of this model produce a kind of strict domination effect in a weighted-constraint system
- Show how this model makes predictions closer to observed typology than a weighted-constraint grammar alone

Introduction

- Weighted-constraint grammars have been criticized for predicting patterns not observed in typology
  - gang effects; e.g. Legendre et al. (2006), but see Pater (2009)
- However, we may want the extra representational power
  - stress windows, "general-case" neutralization
- Additional advantages of weighted constraints:
  - Can use connectionist and statistical methods in learning models
  - Can model variable phenomena (e.g. Maximum Entropy grammar)
    - But see e.g. Riggle (2010) on modeling typological frequency and variation with ranked constraints

Gang effects

- Weighted constraints allow for "gang effects"
- Multiple violations of lower-weighted constraint(s) can cumulatively 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

In This Talk

- Pairing Maximum Entropy grammar with an agent-based, interactive learning model to generate gradient typological predictions
  - Keeps the representational power of the base grammatical model
  - Restricts typological overprediction by assigning low probability to typologically rare or unobserved patterns
- The interactive learning model:
  - avoids variation
  - avoids cumulative patterns (gang effects)
- These effects make this system behave more like a ranked-constraint system (with the same constraints)

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
- 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:
  - 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 (Perceptron update rule; see also Stochastic Gradient Descent, HG-GLA):
- 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

Minimal Working Example

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

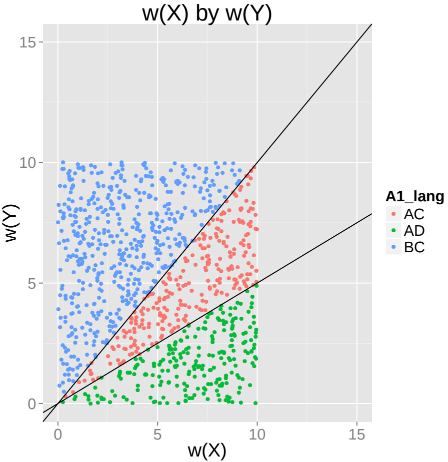
	X	Y
A		-1
B	-1	
D		-2
C	-1	

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

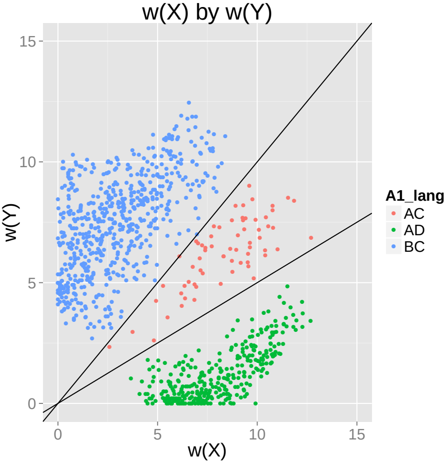
Learning Simulations

- The Interactive learning model was run 1,000 times
- Agents were initialized with random constraint weights sampled from a uniform distribution ranging 0-10
- Agents interacted for 10,000 learning steps
- Baseline prediction is the proportion of possible weights that generate each language type
    - BC: 0.5, AD: 0.25, AC: 0.25

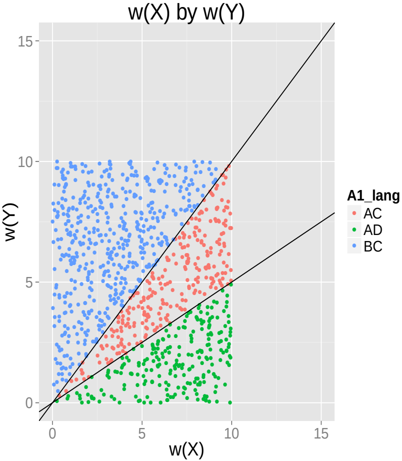
Simulation Start

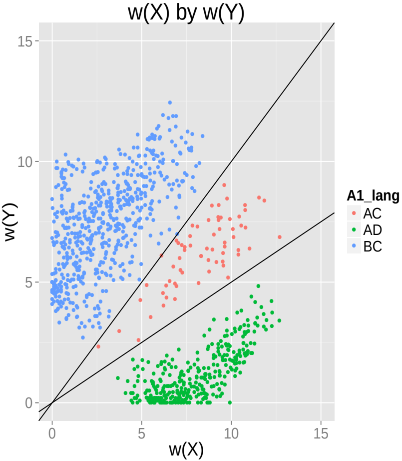


Simulation End

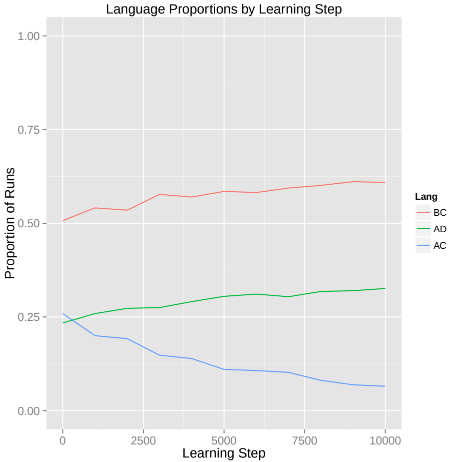


Simulation Results





Simulation Results



Summary

- The interactive learning model:
  - Trended away from variable grammar states
    - Agents moved away from the borders between language types, where there is more variation
  - Trended away from the cumulative pattern
    - Proportion of runs in the gang effect AC pattern decreased as time increased
    - Not representable by a ranking of the same constraints

Stress Window Typology

- The toy example is a mini two-syllable stress window system
- Stress window: stress e.g. a heavy syllable if it falls within a certain distance from the word edge, otherwise, default stress to word edge

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

Stress Window Typology

- The toy example is a mini two-syllable stress window system
- Stress window: stress e.g. a heavy syllable if it falls within a certain distance from the word edge, otherwise, default stress to word edge

/HL/	3 WEIGHT-TO-STRESS	2 ALIGN-R	H
→**H**L		-1	-2
H**L**	-1		-3
/HLL/	WEIGHT-TO-STRESS	ALIGN-R	H
→HL**L**	-1		-3
HLL		-2	-4

Observed stress window data

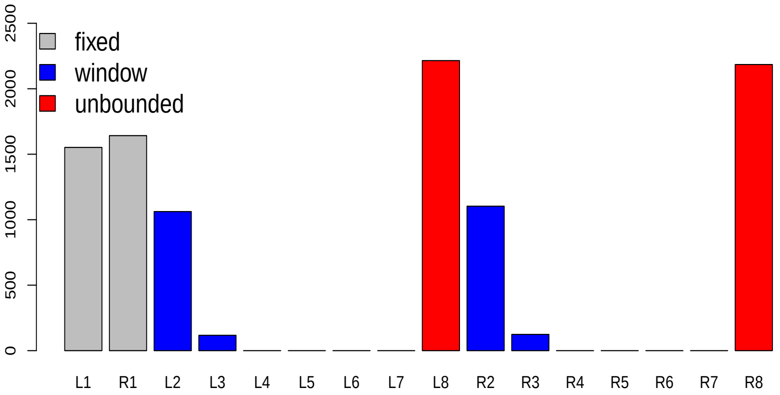
- Counts are adapted from Staubs (2014), Kager (2012), StressTyp (Goedemans, van der Hulst & Visch 1996)

Window Length	Count
2	121
3	39
4 or more	0

Stress Window Simulations

- Constraints include WEIGHT-TO-STRESS, alignment
- Input words are of length 2-8
- Initial agent weights are sampled from a uniform distribution 0-20
- 10,000 runs were performed; runs ended when winning candidate in each tableau had at least 0.95 probability
- Typology includes **fixed**-edge stress, **unbounded** weight-sensitive stress, and stress **window** systems of different lengths

Stress Window Simulation Results



Summary

- Observation: two-syllable windows are more common than three, and larger windows are not attested
- Stress windows could be represented in OT with different constraints, however, this does not (necessarily) predict the observed typological gradient
- Simulation results predict the attested disparity between syllable window sizes
    - Because of the gradient predictions derived from the learning model; the base grammatical model can still theoretically represent any size window

Palatalization Typology

- Palatalization typology: possible contrast patterns between /s/ and /ʃ/ (Carroll 2012)
- Constraints: NO[ʃ], NO[si], IDENT
- With these constraints, 5 possible languages:
  - (44%) Total Neutralization (TN)

    [si], [sa]
  - (37%) Full Contrast (FC)

    [si], [ʃi], [sa], [ʃa]
  - (10.3%) Complementary Distribution (CD)

    [ʃi], [sa]
  - (8.2%) Special-Case Neutralization (SCN)

    [ʃi], [sa], [ʃa]
  - (0.5%) General-Case Neutralization (GCN; gang effect)

    [si], [ʃi], [sa]

General-Case Neutralization (GCN; gang effect)

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

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

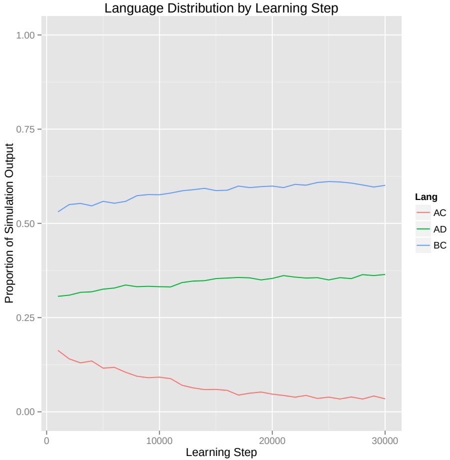
Discussion

- Ranked constraints:
  - No constraint cumulativity
- Weighted constraints can represent everything ranked constraints can, plus:
  - Constraint cumulativity (gang effects)
  - With MaxEnt, can define a probability distribution over outputs
- MaxEnt + Interactive learning model, as outlined in this talk:
  - Has the representational power of MaxEnt
  - Typological predictions trend away from variability and cumulativity, producing a kind of strict domination effect
  - The predictions outlined here are consistent with observed typological facts

Thanks!

Thanks to Gaja Jarosz, Robert Staubs, audiences at UMass, PhoNE, NECPhon, mfm, and everyone here.

Longer toy simulation



Full Stress Window constraint set

- ALIGN-FT-L, ALIGN-FT-R (between foot edge and word edge)
- ALIGN-HEAD-L, ALIGN-HEAD-R (between head foot edge and word edge)
- FTBIN
- *CLASH, *LAPSE
- IAMB, TROCHEE

Simulation Results