Grammatical Agent-Based Modeling of Typology

Coral Hughto      Joe Pater      Robert Staubs

University of Massachusetts Amherst

Workshop on Computation, Learnability and Phonological Theory, GLOW, Paris April 18, 2015
The standard generative approach to typology, grammar, and learning

- A theory of grammar is designed to generate all and only possible languages (typological modeling)
- A theory of learning is designed to find a correct grammar for all languages in the space defined by the grammatical theory
- The learning theory plays no role in the modeling of typology
The ‘iterated learning’ approach to typology, grammar and learning

Pioneered by Luc Steels, Bart de Boer, and Simon Kirby and colleagues (see e.g. Kirby and Hurford 2002), developed in a Bayesian framework by Griffiths and Kalish (2007) and Kirby et al. (2007), see Dediu (2009) for a useful overview and further extensions, Wedel (2007, 2011), Mackie and Mielke (2011) and Rafferty et al. (2013) for related work in phonology.

- Uses agent-based modeling (iterated learning) to study the role of learning and transmission in shaping typology.
- Uses very simple language models, and often highlights emergence of putatively innate features of language.
Grammatical Agent-Based Modeling of Typology: a ‘third way’?

- What is the effect of learning on typology with a relatively richly specified theory of grammar?
- We take no position on the innateness or even universality of the grammatical features, but we do take them as given in the work we discuss here.
Related streams of research:


- Learning and phonological typology – see e.g. Alderete (2008), Boersma and Hamann (2008), Moreton (2008), Heinz (2009) and colleagues, Stanton (2014: AMP, NECPhon)
Overview

We’ll argue that abstracting from learning in typological work may lead to:

- Missed opportunities
- Faulty inferences
**Missed opportunities**

By integrating learning with standard grammatical assumptions, we can capture typological generalizations that escape the grammatical theories working on their own, both in terms of predicting relative frequencies of attested patterns, and cases of (near-)zero attestation.
Faulty inferences

Integrating learning can skew the typologies away from what would be predicted from the grammatical model alone: the output of our ABMs tends away from variation and gang effects, even though we are working with probabilistic weighted constraint grammars.
General constraints and tendencies to generality

Feature geometric nodes simplify spreading or delinking of a whole feature class

- But we can still formalize spreading or delinking for any arbitrary set of features

Padgett’s (2002) feature classes allow a single constraint to target a whole class

- But we can still formalize spreading or delinking for any arbitrary set of features
The HG typology is the same as OT, or a parametric theory.

In all of these frameworks, in terms of the set of systems that our theory can represent, the general constraint is doing no work for us.
The general constraint does make general pattern easier to learn, and it emerges with greater frequency from our ABM.

We also seem to see tendencies to generality in other domains, and these can also be represented with a similar specific/general constraint structure (e.g. lexically specific and regular stress in Pater and Moreton 2012, category specific and consistent syntactic headedness in Pater 2012)
We assume a probabilistic version of HG in which the probability of a candidate is proportional to the exponential of its Harmony (MaxEnt grammar, Goldwater and Johnson 2003, loglinear model in Ernestus and Baayen 2003)

<table>
<thead>
<tr>
<th></th>
<th>No-Coda-Labial</th>
<th>Ident-Labial</th>
<th>No-Coda-Place</th>
<th>Ident-Place</th>
<th>H</th>
<th>exp(H)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>lap</td>
<td>lap</td>
<td>−1</td>
<td>−1</td>
<td>−2</td>
<td>0.14</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>la?</td>
<td>−1</td>
<td>−1</td>
<td>−1</td>
<td>−1</td>
<td>0.37</td>
<td>0.73</td>
<td></td>
</tr>
</tbody>
</table>
Learning: difference between datum produced by teacher and learner’s expectation scaled by a learning rate (below always 0.1), added to current weights

<table>
<thead>
<tr>
<th></th>
<th>No-Coda-Labial</th>
<th>Ident-Labial</th>
<th>No-Coda-Place</th>
<th>Ident-Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>lap</td>
<td>−1</td>
<td></td>
<td>−1</td>
<td></td>
</tr>
<tr>
<td>la?</td>
<td>−1</td>
<td>−1</td>
<td>−1</td>
<td>−1</td>
</tr>
<tr>
<td>T − L</td>
<td>−1</td>
<td>+1</td>
<td>−1</td>
<td>+1</td>
</tr>
</tbody>
</table>

**Gradient ascent:** training data given in batch.

**Stochastic gradient ascent (SGA):** one piece of training data at a time.

**Sampling SGA:** learner expectation also sampled from distribution defined by grammar
It is not safe to assume that a language that is easier learned will be more frequent in a model of typology (Rafferty et al. 2013); amongst other things, it matters what languages become when they are ‘mislearned’

We use interactive learning (Pater 2012): a simple way to generate typologies (see also de Boer’s imitation game)
Interactive learning

- Two agents are started with zero weight on the constraints, and hence equal probability for the two candidates in each tableau.
- In each trial, an agent is picked at random as the teacher, a tableau is picked at random, and learning datum is sampled from the distribution defined by the teacher’s grammar (sampling SGA).
For these simulations, this was done 10,000 times to generate a language; we’ll report results over 100 languages.

There are 4 places of articulation each with a binary choice, so there are $2^4 = 16$ possible combinations across them. Of these, only $2/16$ have consistent (non-)application across places: 0.125 is a baseline probability of consistent application.

Taking the higher probability candidate from each tableau as the output, $74/100 = 0.74$ have consistent application.
100 languages generated without the general constraints produced 7 cases of consistent application (0.07, similar to 0.125 baseline)

- This would be the predicted probability of having four unrelated features pattern together, for example, a process that targets labials along with consonants that are either [+voice], [–continuant] or [+sonorant]
A tendency away from variation can be seen by looking at the probabilities granted to the candidates at the end of simulations.

<table>
<thead>
<tr>
<th></th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run 4</th>
<th>Run 5</th>
<th>Run 6</th>
<th>Run 7</th>
<th>Run 8</th>
<th>Run 9</th>
<th>Run 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-1</td>
<td>0.01</td>
<td>0.90</td>
<td>0.04</td>
<td>0.03</td>
<td>0.68</td>
<td>0.91</td>
<td>0.03</td>
<td>0.79</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>P-2</td>
<td>0.01</td>
<td>0.96</td>
<td>0.03</td>
<td>0.03</td>
<td>0.95</td>
<td>0.88</td>
<td>0.04</td>
<td>0.92</td>
<td>0.10</td>
<td>0.49</td>
</tr>
<tr>
<td>P-3</td>
<td>0.01</td>
<td>0.89</td>
<td>0.06</td>
<td>0.05</td>
<td>0.65</td>
<td>0.97</td>
<td>0.05</td>
<td>0.89</td>
<td>0.13</td>
<td>0.75</td>
</tr>
<tr>
<td>P-4</td>
<td>0.01</td>
<td>0.94</td>
<td>0.07</td>
<td>0.06</td>
<td>0.96</td>
<td>0.94</td>
<td>0.38</td>
<td>0.97</td>
<td>0.02</td>
<td>0.42</td>
</tr>
</tbody>
</table>

*Probability of unfaithful candidate at end of simulation, at each place of articulation averaged over agents*
Summing up:

- Adding learning to our model of typology allows a general constraint to do the work we might expect it should: it produces a tendency toward generality → avoiding a missed opportunity
- The output of our ABM also shows a tendency toward deterministic outcomes even though our grammar model is probabilistic → avoiding a faulty inference

Looking ahead:

- Why the tendency to deterministic outcomes?
- Learning and deterministic outcomes in skewing away from a gang effect
- Stress window typology and gradient predictions
Tendency to deterministic outcomes

Why do the agents’ grammars tend towards deterministic outcomes?

1. Weight changes can push the agents towards either more or less deterministic states.
2. As the agents drift into deterministic grammars:
   1. They change less and less: the agents are less likely to disagree, triggering an update
   2. The effective change to probability from a change in weights shrinks (because of MaxEnt)
3. The system tends to stay in deterministic states once it reaches them
The network structure of this agent-based model is distinct from those in Kirby and Hurford (2002) and Griffiths and Kalish (2007), besides the fact that there are only two agents (at a time) in both.

Their models are purely generational: a designated learner learns for some number of trials from a designated teacher, and then becomes the teacher for the next generation.

This model is purely interactive: agents always have the same probability of being the teacher or learner.

Reality lies between these two poles; in any case the purely generational model also produces the deterministic tendency (Dediu 2009: 555, Hughto et al. 2014)
Learning, deterministic outcomes, and gang effects

A simple gang effect:
A beats B if \( w(X) > w(Y) \), and C beats D if \( 2w(Y) > w(X) \)

<table>
<thead>
<tr>
<th>Two constraint gang effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>
Gang effects might in general be hard to learn because they require relatively precise weight ratios (Prince 1993/2007). When we start weights at zero and use gradient ascent, learning is much slower for the gang effect in our two constraint example.
A margin of separation of 3 between the Harmony scores of the two candidates grants just over .95 probability to one of them.

- The lowest weights for the non-gang effect cases:
  - **AD**: X = 3, Y = 0
  - **BC**: X = 0, Y = 3

- The lowest weights for the gang effect:
  - **AC**: X = 9, Y = 6
The gang effect also occurs in a relatively small portion of the weight space generating nearly deterministic outcomes.

1000 runs of a simulation where the two agents, starting at zero constraint weights, exchange data until there is at least 95% probability on one candidate in each tableau.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-1</td>
</tr>
<tr>
<td>B</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>-1</td>
</tr>
<tr>
<td>D</td>
<td>-2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, D</td>
<td>312</td>
</tr>
<tr>
<td>B, C</td>
<td>688</td>
</tr>
<tr>
<td>A, C</td>
<td>0</td>
</tr>
</tbody>
</table>
Simulation begun with initial constraint weights randomly sampled from uniform distribution 0-10
Stress windows

Potential problem for HG: it can generate a stress window of any size (Legendre, Sorace and Smolensky 2006, Pater 2009)

An unbounded trade-off

<table>
<thead>
<tr>
<th>/banσₙ.ta/</th>
<th>WEIGHT-TO-STRESS</th>
<th>MAINSTRESSRIGHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ban.σₙ.tá</td>
<td>−1</td>
<td></td>
</tr>
<tr>
<td>bán.σₙ.ta</td>
<td>−1–n</td>
<td></td>
</tr>
</tbody>
</table>
Only 3-syllable windows exist, not 4-syllable or larger

3-syllable windows are much rarer than 2 (Staubs 2014b: 89)

<table>
<thead>
<tr>
<th>Window type</th>
<th>#</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final two syllables</td>
<td>82</td>
<td>e.g. Yapese (Jensen et al. 1977)</td>
</tr>
<tr>
<td>Final three syllables</td>
<td>38</td>
<td>e.g. Pirahã (Everett and Everett, 1984)</td>
</tr>
<tr>
<td>Initial two syllables</td>
<td>39</td>
<td>e.g. Malayalam (Asher and Kumari, 1997)</td>
</tr>
<tr>
<td>Initial three syllables</td>
<td>1</td>
<td>e.g. Comanche (Smalley, 1953)</td>
</tr>
</tbody>
</table>

Kager (2012)
- words 2–8 syllables
- 10,000 runs
- 0.95 stopping criterion

- left and right alignment
- initial weights at zero

![Graph depicting stressed windows and alignment with fixed and unbounded windows.](image-url)
- words 2–8 syllables
- 10,000 runs
- 0.95 stopping criterion
- left and right alignment

- initial weights sampled from uniform distribution 0–20
weights sampled from uniform distribution 0–20

no learning

no 0.95 criterion
Results show gradient differences in frequency between language types, just like the real typology.

→ no missed opportunity

Categorical impossibility replaceable with near-zero predicted frequency.

→ no problem with HG in this setting
HG vs. OT

Ranking vs. weighting: our model uses a weighted constraint grammar, but shows skews away from gang effects

- Much remains to be done in terms of studying the generality of this result, but the potential consequences are striking, since the grammatical model alone might overgenerate, with the full model remaining sufficiently restrictive.
Conclusions

- In generative linguistics, we typically make relatively direct inferences about the structure of UG from typology.
- These results suggest that abstracting from learning may not be entirely safe (deterministic outcomes, gang effects) and may result in missed opportunities (generality from general constraints, accounts of tendencies).
Thank you!

This material is based upon work supported by the National Science Foundation under Grant No. S12100000211 to the second author, Grants BCS-0813829 and BCS-424077 to the University of Massachusetts Amherst, and by the city of Paris under a Research in Paris fellowship to the first author.

We would like to thank Lucien Carroll, Joaquim Brandao de Carvalho, Ewan Dunbar, Giorgio Magri and Elliott Moreton for discussion, as well as John McCarthy and other members of the phonology grant group at UMass.