Linguistics Colloqium
UMass-Amherst

# Syntagmatic simplicity bias in human and artificial learners

February 20, 2009

*Elliott Moreton*[1]
*University of North Carolina, Chapel Hill*

## 1 Introduction

(1) Phonotactic learning:

a. In nature: Implicit knowledge of L1 phonotactics acquired starting in first year of life; effects detectable in adults in a wide range of tasks.

b. In the lab: Knowledge of phonotactic patterns acquired rapidly from training, detectable in wide range of tasks.

c. Subject to *inductive bias*, in the sense that the size of the phonotactic effect depends partly on the content of the pattern and not just on how consistently it is instantiated in the training data. (Artificial: Schane et al. (1974); Wilson (2003); Carpenter (2005); Wilson (2006). L1: Zimmer (1969); Hsieh (1976); Becker et al. (2007); Hayes et al. (2008). L2: Kager et al. (2008).)

(2) Today's topic: "syntagmatic simplicity bias", a learning advantage enjoyed by single-feature over two-feature dependencies. E.g.,

$$[+\mathit{voice}] \Longleftrightarrow [+\mathit{voice}] \text{ is easier than } [+\mathit{high}] \Longleftrightarrow [+\mathit{voice}]$$

Of interest to linguists because

a. *Typologically effective*—typology is skewed by it in ways that can't be attributed to differences in the size of the phonetic precursors available for phonologization (Moreton, 2008).

b. *Formal* rather than substantive; it has to do with the features themselves rather than their real-world interpretation (Hale and Reiss, 2000).

(3) Preview:

§2 "Syntagmatic simplicity bias": Humans learn single-feature dependencies better than two-feature ones, regardless of what the features actually are.

§3 Connection to paradigmatic simplicity bias (natural classes): The patterns that are learned faster are supported by multiple overlapping constraints ("wholesale" vs. "retail" constraints).

§4 Current models of constraint induction induce only retail constraints. Here's a suggestion for improvement; preliminary results are encouraging.

§5 Summing-up and future directions.

[1] I am indebted to many people for discussion of the ideas and facts presented here, especially Adam Albright, Joe Pater, Jen Smith, Anne-Michelle Tessier, and Alan Yu. Work in §4 is part of a collaboration with Joe Pater and Michael Becker. Thanks are also due to Abby Spears of UNC for subject-running and data-transcribing, and to Chris Wiesen of UNC's Odum Institute for statistical advice. Supported in part by a grant from the University Research Council (UNC). Remaining errors are mine. Email may be addressed to `moreton@unc.edu`.

## 2 Syntagmatic simplicity in human learners

(4) Outline: In a $C_1V_1C_2V_2$ stimulus,

§2.1 Height agreement (“HH”) and voice agreement (“VV”) are learned better than a dependency between $V_1$ height and $C_2$ voice (“HV”). Some of that may be due to an advantage for vowel-vowel dependencies over vowel-consonant ones. . .

§2.2 . . . but not all of it, since height-backness (“HB”) and place-voice (“PV”) dependencies are not learned any better than HV.

⇒ Doesn’t depend on what the features actually are, just whether they are the same feature.

### 2.1 Experiments 1 and 2: height-height, voice-voice > height-voice.

(5) Stimuli: MBROLA-synthesized $C_1V_1C_2V_2$ words with inventory /t k d g/ /i u æ ɔ/. Two patterns:

a. “HH pattern”: Vowels agree in height.

b. “HV pattern”: $V_1$ high iff $C_2$ voiced.

(6) Experimental paradigm (based on Moreton (2008, Exps. 1 and 2)):

a. *Study Phase*: Listen to pattern-conforming words through headphones, repeat into microphone. 32 words × 4 repetitions, randomized in blocks.

| Pattern conformity | | Training condition | |
|---|---|---|---|
| HH | HV | HH | HV |
| + | + | 16 | 16 |
| + | − | 16 | |
| − | + | | 16 |
| − | − | | |

b. *Test Phase*: Listen to pairs of new words, choose the one that you think is “a word of the language you studied”. 32 pairs in two counterbalanced blocks of 16, random orders in block and pair. Each pair pits one pattern-conforming item against one pattern-nonconforming item:

| Pattern conformity | | | | | Studied pattern | |
|---|---|---|---|---|---|---|
| HH | HV | | HH | HV | HH | HV |
| + | + | *vs.* | − | − | 16 | 16 |
| + | − | *vs.* | − | + | 16 | 16 |

(7) Properties of this design:

a. For half of the Test pairs, the correct response depends on the Study pattern; for the other half, it does not. Allows effects of learning to be separated from those of pre-existing preferences.

b. Does *not* test generalization to new vowels or new combinations of vowels (i.e., does not distinguish between learning vowel harmony and learning a list of vowel-vowel sequences).

(8) Participants: 18 native speakers of American English. None had studied or otherwise learned a language with vowel harmony. One explicitly noticed pattern (post-experiment questionnaire) and was replaced.

(9) Results of Experiment 1[2].

a. Performance in HV condition was not distinguishable from chance.

b. Participants in HH condition nearly doubled their odds of a correct response, in both the first and second half of the Test phase.

| Coefficient | Estimate | SE | $z$ | $Pr(>\mid z\mid)$ | |
|---|---|---|---|---|---|
| *(Intercept)* | 0.27419 | 0.19609 | 1.39830 | 0.162024 | |
| *Studied HH* | 0.71606 | 0.27884 | 2.56804 | 0.010228 | * |
| $V_1 = V_2$ | –0.25962 | 0.20536 | –1.26420 | 0.206160 | |
| *2nd half* | –0.27877 | 0.24170 | –1.15339 | 0.248750 | |
| *Studied HH* × *2nd half* | –0.05977 | 0.35390 | –0.16889 | 0.865882 | |
| *HH-nonconforming* | 0.10146 | 0.13140 | 0.77217 | 0.440015 | |
| *1st in pair* | 0.46502 | 0.17679 | 2.63042 | 0.008528 | ** |

(10) ⇒ HH pattern learned better than HV.

(11) Exp. 2 was like Exp. 1, except that a voice-voice pattern replaced the height-height one:

a. "VV pattern": $C_1$ and $C_2$ agree in voicing.

b. "HV pattern": $V_1$ high iff $C_2$ voiced, as in Exp. 2.

(12) Results of Experiment 2.

a. Participants in the HV condition were again at or near chance.

b. Those who studied the VV pattern doubled their odds of a correct response. The effect did not diminish significantly over the course of testing.

| Coefficient | Estimate | SE | $z$ | $Pr(>\mid z\mid)$ | |
|---|---|---|---|---|---|
| *(Intercept)* | 0.157994 | 0.225038 | 0.70208 | 0.482631 | |
| *Studied VV* | 0.736347 | 0.309506 | 2.37911 | 0.017355 | * |
| $C_1 = C_2$ | –0.480821 | 0.199329 | –2.41219 | 0.015857 | * |
| *2nd half* | 0.022876 | 0.246716 | 0.09272 | 0.926125 | |
| *StudiedVV* × *2nd half* | –0.540257 | 0.348545 | –1.55004 | 0.121133 | |
| *VV-nonconforming* | 0.271081 | 0.132973 | 2.03862 | 0.041488 | * |
| *1st in pair* | 0.468015 | 0.174332 | 2.68462 | 0.007261 | ** |

(13) ⇒ VV learned better than HV also.

[2] Analyzed by mixed-effects logistic regression with Participant as a random effect. The independent variables were chosen as follows: Each of the 6 experiments in this series was modelled using a larger set of terms. The models were then reduced by backwards elimination. Any term which could not be eliminated from at least *one* of the 6 models was retained (mutatis mutandis) in the analysis of all of them.

## 2.2 Experiments 3 and 4: Height-backness and place-voice ≈ height-voice

(14) Maybe HH and VV beat HV only because within-tier dependencies (vowel-to-vowel or consonant-to-consonant) are easier than cross-tier ones (vowel-to-consonant). Let's see:

(15) Exp. 3 was like Exp. 1, except that a height-*backness* pattern replaced the height-height one:

a. "HB pattern": $V_1$ high iff $V_2$ back

b. "HV pattern": $V_1$ high iff $C_2$ voiced.

(16) Results of Experiment 3.

a. Once again, those who studied the HV pattern performed at chance.

b. Those who studied the HB pattern did *marginally* better, but the difference did not reach the conventional 5% criterion, and disappeared entirely in the second half of the Test phase.

| Coefficient | Estimate | SE | $z$ | $Pr(>\mid z \mid)$ | |
|---|---:|---:|---:|---:|---|
| *(Intercept)* | –0.099234 | 0.197518 | –0.50241 | 0.61538 | |
| *Studied HB* | 0.495776 | 0.284555 | 1.74228 | 0.08146 | . |
| *2nd half* | 0.104045 | 0.239182 | 0.43500 | 0.66356 | |
| *Studied HB* × *2nd half* | –0.583042 | 0.343219 | –1.69875 | 0.08937 | . |
| *HB-nonconforming* | –0.115660 | 0.119626 | –0.96685 | 0.33362 | |
| *1st in pair* | 0.455936 | 0.171834 | 2.65335 | 0.00797 | ** |

(17) Exp. 4 was similar, but used a place-voice dependency:

a. "PV pattern": $C_1$ velar iff $V_2$ voiced

b. "HV pattern": $V_1$ high iff $C_2$ voiced (as in Exp. 2).

(18) Results of Experiment 4.

a. Yet again, studying the HV pattern led to near-chance performance.

b. Those who studied the PV pattern did *marginally* better, but only in the first half of the Test phase.

| Coefficient | Estimate | SE | $z$ | $Pr(>\mid z \mid)$ | |
|---|---:|---:|---:|---:|---|
| *(Intercept)* | –0.21116 | 0.19910 | –1.06054 | 0.288901 | |
| *Studied PV* | 0.48402 | 0.28433 | 1.70231 | 0.088697 | . |
| *2nd half* | 0.16938 | 0.24039 | 0.70461 | 0.481052 | |
| *Studied PV* × *2nd half* | –0.42830 | 0.33915 | –1.26286 | 0.206640 | |
| *PV-nonconforming* | 0.21102 | 0.12021 | 1.75546 | 0.079181 | . |
| *1st in pair* | 0.49330 | 0.16986 | 2.90418 | 0.003682 | ** |

(19) Exps. 3 and 4 show that when other factors are controlled, one-feature dependencies are learned better than two-feature dependencies.

a. The results of 1 and 2 resemble each other, as do those of 3 and 4. Apparently the content of the features doesn't matter, only their formal arrangement.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Coefficient | HH | VV | HB | PV |
| *(Intercept)* | 0.274 | 0.157 | –0.099 | –0.211 |
| *Studied XY* | 0.716 * | 0.736 * | 0.495 . | 0.484 . |
| $V_1 = V_2$ *or* $C_1 = C_2$ | –0.259 | –0.480 * | — | — |
| *2nd half* | –0.278 | 0.022 | 0.104 | 0.169 |
| *Studied XY* × *2nd half* | –0.059 | –0.540 | –0.583 | –0.428 |
| *XY-nonconforming* | 0.101 | 0.271 * | –0.115 | 0.211 . |
| *1st in pair* | 0.465 ** | 0.468 ** | 0.455 ** | 0.493 ** |

## 3 Simplicity, multiplicity, and generality

(20) ⇒ Learning a syntagmatic dependency between two feature instances is facilitated when they are both instances of the *same* feature.

a. Corroborates Wilson (2003)'s finding that a [nasal]-[nasal] dependency is learned better than a [nasal]-[Dorsal] dependency.

b. It does not seem to matter whether the dependency has a robust phonetic precursor or not—i.e., the real-world interpretation of the feature instances is irrelevant; all that matters is that they are the same.

(21) "Paradigmatic simplicity bias": Experiments of Saffran and Thiessen (2003).

a. 9-month olds, familiarized on a list of pattern-conforming nonsense words, then exposed to two pattern-conforming and two pattern-nonconforming words, then tested to see how long they liked to listen to each of those 4 words.

b. Exp. 2 (SIMPLE): Words were CVCCVC, where CVC = [ptk]V[bdg] (or the reverse, for half of the participants). Result: Liked the non-conforming words more. ⇒ Learned the pattern.

c. Exp. 3 (COMPLEX): Like Exp. 2, but CVC = [p**d**k]V[p**t**g] (or vice versa). Result: No difference in listening time.

(22) There's already a proposal about paradigmatic simplicity bias (Pater, 2008; Pater et al., 2008).

a. Induction of two-feature constraints yields *multiple overlapping constraints* supporting the [ptk]-vs.-[bdg] pattern (e.g., "Be voiceless", which overlaps with "Be voiceless and labial", "Be voiceless and coronal", "Be voiceless and dorsal").

b. Induction yields *isolated* constraints supporting the [pdk]-vs.-[btg] pattern (e.g., "Be voiceless and labial", "Be voiced and coronal", "Be voiceless and dorsal") which conflict with more general constraints (e.g., "Be voiceless" conflicts with "Be voiced and coronal").

c. When the induced constraints are given to a Maximum Entropy learner, it learns (goes from chance to any given criterion) faster in the [ptk] case than the [bdg] one.

(23) Can we extend this explanation to paradigmatic simplicity bias? Two basic ways to implement if-and-only-if patterns:

a. Two “retail” constraints share the work: $*+-$ and $*-+$.

b. A single “wholesale” constraint does it all: $*\{+-,-+\}$ (i.e., an Agree constraint).

What if HV pattern is supported only by retail constraints, whereas HH and VV are supported by both wholesale and retail constraints?

(24) Depends on learning algorithm. Stick with the Maximum Entropy algorithm as formulated by Jäger (to appear), for compatibility with Pater et al. (2008) and Hayes and Wilson (2008), and because the math is easier than with the Gradual Learning Algorithm (Boersma, 1998; Boersma and Hayes, 2001). Assumptions:

a. A “pure phonotactic learner” in the sense of Hayes (2004); its goal is to distinguish well-formed from ill-formed surface strings.

b. Only one input, /$i$/; candidate outputs are $++, +-, -+, --$.

c. No faithfulness constraints; only markedness constraints $c_1, \ldots, c_N$. Constraint $c_j$ assigns $c_j(o)$ violation marks to candidate [$o$].

d. Constraints are not ranked, but weighted with weights $w_1, \ldots, w_N$, which can be positive or negative.

e. For a given set of weights, the probability of a given output [$o$] is proportional to $\exp(\sum_j w_j c_j(o))$.

f. All constraints have initial weight 0.

g. Training data is sampled from $\{++, --\}$ with uniform probability.

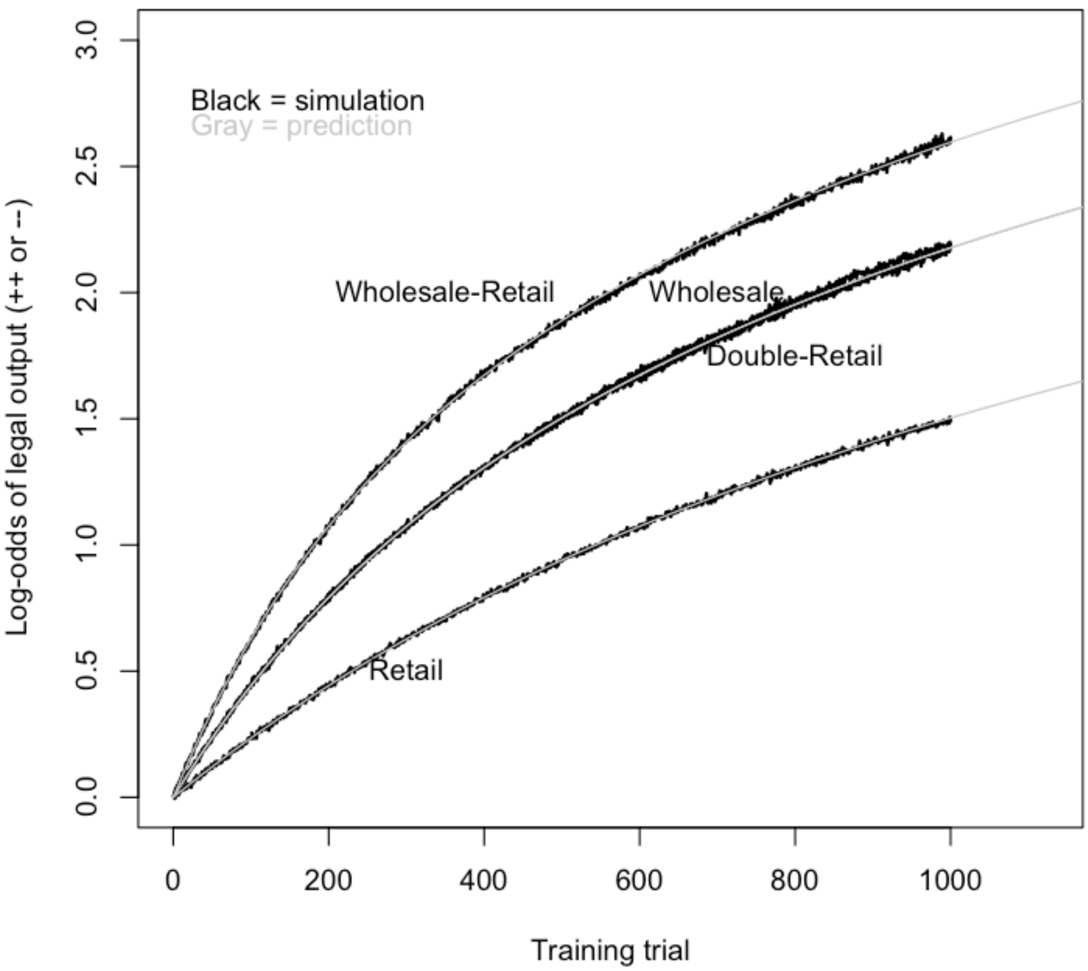
(25) We start with the Retail learner, which has only the two constraints $*[+-]$ and $*[-+]$. If we set any success criterion $\Omega$ in terms of the odds that the learner will map /$i$/ to one of $\{++, --\}$, we can show (details omitted) that the time $t$ required to reach that criterion is, on average,

$$t = \frac{2}{\eta}(\Omega + \ln(\Omega) + 1)$$

where $\eta$ is the learning rate (the nudge each constraint gets when the learner makes a relevant error).

(26) A Wholesale+Retail learner reaches the same criterion after only $t/3$ training trials (simulations in Praat, Boersma & Weenink 2005) :

| Retail | Double Retail | Wholesale | Wholesale+Retail |
|---|---|---|---|
| $t$ | $t/2$ | $t/2$ | $t/3$ |
| $*[+-]$, $*[-+]$ | $*[+-]$, $*[-+]$ | | $*[+-]$, $*[-+]$ |
| | $*[+-]$, $*[-+]$ | | |
| | | $*\{[+-],[-+]\}$ | $*\{[+-],[-+]\}$ |



(27) The Wholesale-Retail learner's advantage is due to two factors:

a. *Constraint multiplicity*: All of the positive candidates are supported by two constraints rather than one. In fact, we can get a twofold increase in the Retail learner's speed by simply duplicating the Retail constraints in the Double-Retail learner.

b. *Constraint generality*: The learner promotes the Wholesale constraint in response to two kinds of errors rather than just one: Training on [++] teaches the learner about [−−], and vice versa. We can double the Retail learner's speed by merging its two constraints to make the Wholesale-only learner.

(28) Yields new question: Why do the HH and VV patterns have wholesale constraints, but not the HV pattern?

## 4 Inducing subtree constraints

(29) Proposed answer: Constraints are induced according to a schema which allows iff relations between two instances of same feature, but not between two instances of different features.

a. Iff constraints represented using variables, e.g., Agree-[high] is $*[\alpha\, high] \ldots [-\alpha\, high]$.

b. Variables are relativized to specific features: $*[\alpha\, high] \ldots [\alpha\, voice]$ (the would-be wholesale HV constraint) is interpreted as $*[\alpha\, high] \ldots [\beta\, voice]$

(30) Why do we need constraint induction?

a. “Crazy rules” (Bach and Harms, 1972) ⇒ induce constraints from *phonological* data.

b. “Inductive grounding” (Hayes, 1999) ⇒ induce constraints from *phonetic* data.

(31) Current constraint inducers won’t help:

a. No syntagmatic feature variables—can’t express wholesale iff constraints like Agree or OCP (Gildea and Jurafsky, 1995; Albright and Hayes, 2002; Heinz, 2007; Hayes and Wilson, 2008)

b. Arbitrary limits on constraint complexity in order to make exhaustive search feasible, leading to . . .

c. . . . difficulties with non-adjacent dependencies (see discussion in Hayes and Wilson (2008, 6.2))

d. Poor integration of features and prosodic structure; positional constraints aren’t supported Hayes and Wilson (2008).

(32) We need to liberalize the constraint schema, without making the space unsearchable. Here is one attempt:

a. §4.1 describes a constraint schema, the Subtree Constraint schema, which provides for syntagmatic variables, constraints of arbitrary complexity, and non-adjacent dependencies.

b. §4.2 discusses an algorithm for inducing Subtree constraints from phonological data.

c. §4.3 presents some (preliminary!) simulation results showing how the Subtree schema and inducer apply to the HH>HV case.

## 4.1 The Subtree Constraint Schema

(33) *Idea*: Represent constraints by representing a locus of violation (or locus of satisfaction). ⇒ Constraints are *subtrees* of representations, and every representation is itself a constraint (Burzio, 1999).

(34) Two basic kinds of node:

a. Features

(i) Binary: $[+\, cont]$, $[-\, nas]$, etc. Have no dependents.

(ii) Unary: $[Place]$, etc. Have named, unordered dependents.

(iii) A feature tree rooted at $A$ matches once in another one rooted at $B$ if and only if $A$ matches in $B$, and all of $A$’s dependents match in their namesakes in $B$.

$$\begin{array}{l}[Place]\\ \quad [Cor]\\ \qquad [+ant]\end{array} \text{ matches once in } \begin{array}{l}[Place]\\ \quad [Cor]\\ \qquad [+ant]\\ \qquad [+dist]\end{array} \text{, but not in } \begin{array}{l}[Place]\\ \quad [Lab]\\ \quad [Cor]\end{array}$$

b. Prosodic categories (“PrCats”)

   (i) Have nameless, ordered dependents.

   (ii) Of which one may be designated the head.

   (iii) May be unanchored, left-anchored, right-anchored, or left- and right-anchored, forcing the match to start or end with the initial or final dependent.

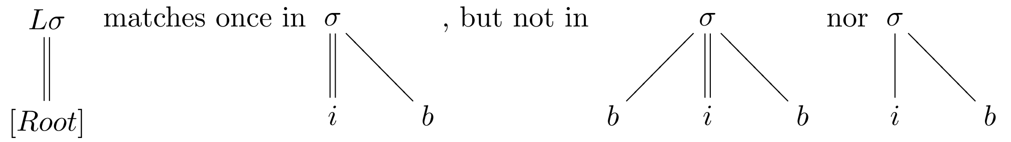
   (iv) A tree rooted at PrCat *A* matches in another one rooted at *B* one time for every way to match every dependent of *A* in a dependent of *B*, preserving order and adjacency, subject to some requirements:

      i. Anchoring

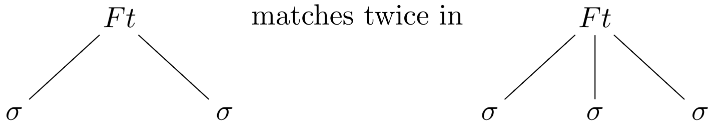
      ii. Heads only match in heads

      iii. *A* must itself match in *B* (both have to be the same prosodic type, and any anchors set in *A* must also be set in *B*).

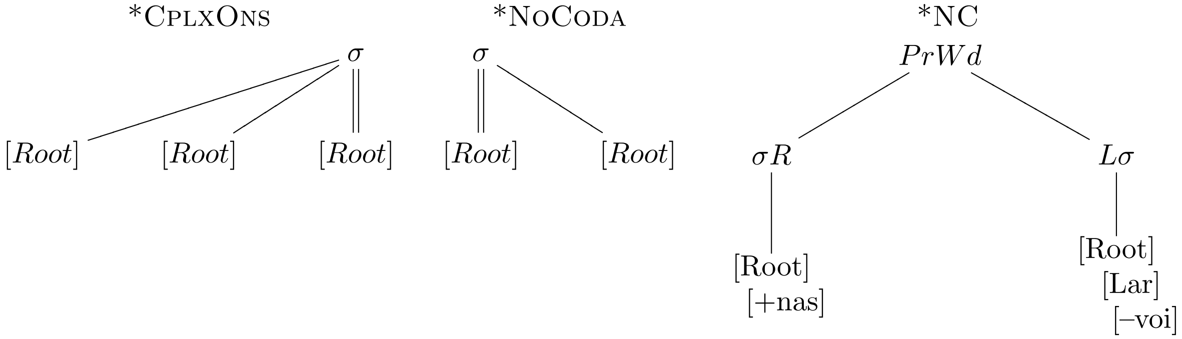
   (v) Here’s Onset, à la (Smith, 2006); letters abbreviate big feature trees rooted at [*Root*]:



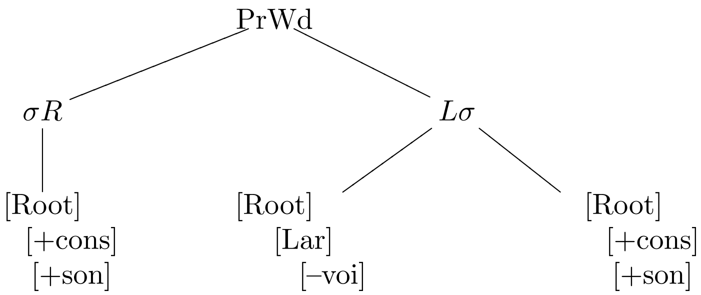
(35) The number of matches returned by a constraint is the number of different ways to match the constraint in the representation. E.g.,



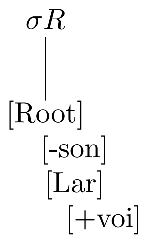
(36) This is a fairly flexible schema, even without variables. Here are a few familiar constraints (negative constraints: “*”, positive constraints: “+”:



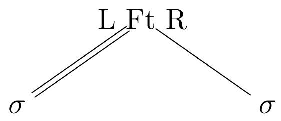
*Intervocalic voicing



*Final obst. devoicing



+FtForm=Trochee



(37) New term: The *genus* of a PrCat node or feature node is the category of which that node is an instance. Examples of genera: PrWd, syllable, Place, constricted glottis, etc.

(38) Variables

a. A variable is itself a kind of node. There is a variable genus for every non-variable genus, e.g., $[\alpha\,_{\mathrm{Syll}}]$ (PrCat), $[\beta\, \mathit{Place}]$ (unary feature), or $[\gamma\, \mathit{voi}]$ (binary feature).

b. The binary ones can have coefficients, e.g., $[-\gamma\, \mathit{voi}]$.

c. Variables have no dependents.

d. A successful match of a constraint in a representation can only occur if

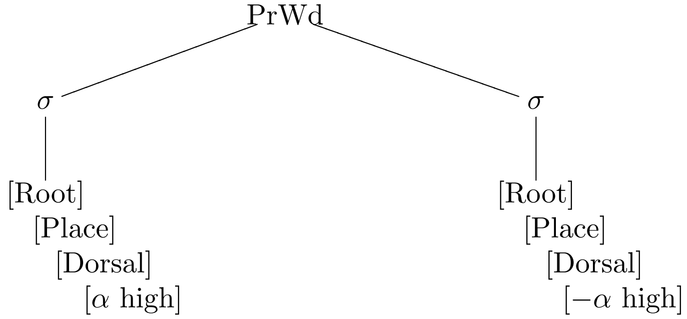
(i) **Every instance of a variable is matched to a node of the same genus**.

(ii) All instances of a given PrCat variable or unary-feature variable in the constraint are matched to identical subtrees of the representation.

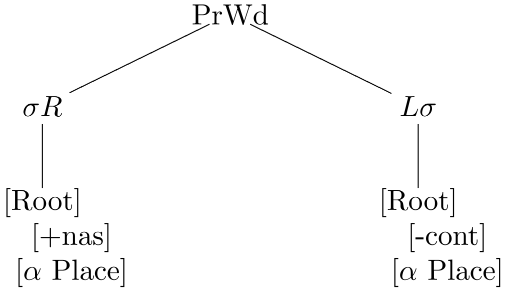
(iii) All instances of a given binary-feature variable are matched so that the variable coefficients times the representation's coefficients are the same (e.g., $[\alpha\, \mathit{back}] \ldots [-\alpha\, \mathit{back}]$ matches in $[+\mathit{back}] \ldots [-\mathit{back}]$ or in $[-\mathit{back}] \ldots [+\mathit{back}]$.

(39) Some familiar constraints with variables:

*Agree-[high]



+Nasal place assimilation



(40) Advantages:

a. can represent lots of different constraint types

b. implements syntagmatic variables ⇒ wholesale if-and-only-if constraints

c. handles long-distance dependencies gracefully

d. sets up the constraint set for faster learning of HH and VV patterns than of HV pattern.

## 4.2 Supervised learning of subtree constraints

(41) But is the space of possible constraints efficiently searchable? We can't search it exhaustively, because it is too big (infinite), but perhaps we can do so non-exhaustively. Here is one attempt.

(42) Non-exhaustive search idea: Use an evolutionary algorithm, so that the learner only has to keep a few schemas in mind at a time, and keeps improving them through small variations (Eiben and Smith, 2003).

a. Population size is fixed.

b. Initial constraint set is the set of training data, since representations are constraints. Similar to Albright and Hayes (2002) Minimal Generalization Learner—start with the data and simplify it.

c. For each constraint, calculate fitness (absolute value of difference between mean number of matches per item in negative and positive data).

d. There is one birth per training cycle. Opportunities to reproduce asexually are raffled off using "fitness-proportional selection" (the fitness of this constraint divided by that of all constraints).

e. The new-born constraint is a recursively mutated copy of its parent, i.e., dependents are copied with mutation.

f. The new mutant joins the population (displacing the least-fit old constraint) iff

(i) It is at least as fit as the least-fit old constraint, and

(ii) It is not an exact duplicate of any old constraint.

g. The algorithm ends after a set number of births.

(43) Goal of the learner: Given positive and negative training data (identified as such) and a time limit, find the $N$ "fittest" constraints, where a "fit" constraint is one which matches more often in the average positive item than negative, or vice versa.

a. All induction is done before any ranking, as in Pater et al. (2008). Not interleaved with ranking, as in Hayes and Wilson (2008).

b. Not necessarily the best criterion for individual constraints—ignores, e.g., within-category variability.

## 4.3 Application to HH>HV

(44) How long does it take the Subtree learner to find the wholesale and retail constraints in a simulation of Exp. 1?

(45) Used the following representational scheme, a slight simplification of a generic one that I took from a phonology textbook(Gussenhoven and Jacobs, 2005). (Missing are: feet, moras,

$[Rad]$, $[\pm constricted\,glottis]$, and $[\pm strident]$.)

- $PrWd$
  - $\sigma$
    - $[Root]$
      - $[\pm cons], [\pm son], [\pm approx], [\pm cont], [\pm nas], [\pm lat]$
      - $[Lar]$
        - $[\pm voice]$
        - $[\pm spread\,glottis]$
      - $[Place]$
        - $[Lab]$
          - $[\pm round]$
        - $[Cor]$
          - $[\pm ant]$
          - $[\pm dist]$
        - $[Dor]$
          - $[\pm high]$
          - $[\pm low]$
          - $[\pm back]$

(46) Simplifications:

a. Training data was from a reduced version of Experiment 1 (HH vs. HV), made by removing the initial unbalanced $C_1$. This yielded one of each of the 64 possible $V_1C_2V_2$ sequences, 32 positive and 32 negative items

b. Turned off prosodic variables.

c. Banned constraints with different variables of the same genus (e.g., $[\alpha\, Dor][\beta\, Dor]$), because I haven't figured out a reasonable mutation scheme.

(47) Results (still skimpy):

| | | Learning trials required to find first | | |
|---|---|---|---|---|
| Parameter set | Run | HV retail | HH retail | HH wholesale |
| I | a. | 31100 | 36500 | 53100 |
| | b. | 100300 | 18700 | 176200 |
| | c. | 60300 | 28000 | 11300 |
| II | a. | 4100 | 10700 | 69900 |
| | b. | 5400 | 5000 | 189600 |
| | c. | 59200 | 1400 | 83100 |

(48) ⇒ Search space isn't *prima facie* unsearchably large.

(49) Main known shortcomings:

a. Needs negative as well as positive data. Possible ways out:

(i) Use English lexicon as negative data.

(ii) Make own negative data by mutating the positive data. (Eisner, "contrastive estimation"?)

b. No science behind parameters (population size, mutation probabilities, etc.)

c. Subtree schema can't *ignore* prosodic structure

d. Doesn't implement negation of non-binary variables, which forces constraints like Place Assimilation to be stated positively rather than negatively. (Probably technical rather than conceptual problem.)

e. Constraints become less diverse over time.

## 5 Conclusions

(50) Main points:

a. Humans learn single-feature dependencies faster than two-feature ones, regardless of real-world interpretation of features ("syntagmatic simplicity bias").

b. Connectable to *paradigmatic* simplicity bias (for featurally-systematic over featurally-arbitrary classes) via effects of constraint multiplicity and generality ("wholesale" and "retail" constraints) on incremental learning of constraint weightings.

c. Current models of constraint induction (from phonological or phonetic data) don't induce wholesale constraints (because their constraint schemas can't represent them), and so can't capture syntagmatic simplicity bias.

d. Subtree Constraint Schema (constraint = representational subtree matching a locus of violation or satisfaction) can represent wholesale/retail distinction, but price is infinite search space.

e. Space may be searchable using non-exhaustive search technique (evolutionary algorithm).

f. Simplicity *emerges* from the way the acquisition mechanisms work, rather than being imposed from outside as a grammar-selection criterion (Hale and Reiss, 2000, fn. 8, p. 164).

Where could this all lead?

(51) Explanation schema:

a. Constraint schema defines constraint space.

b. Induction algorithm plus data (phonetic or phonological) searches space, gets constraints.

c. Ranking algorithm plus constraints and data yields grammar.

(52) ⇒ Explanatory focus shifts to

a. Constraint schemas, generation, and testing (Boersma, 1998; Hayes, 1999; Smith, 2002, 2004; Boersma and Pater, 2007).

b. If constraints are representations, convergence of constraint schemas with representational schemas like Feature Geometry.

c. If space is searched by mutation and selection, considerations of mutation algorithm (connection to gradualism in Harmonic Serialism?)

(53) Example: ways to use constraint multiplicity and generality, plus induction, to account for bias:

a. *Constraint synonymy.* Schema sometimes produces formally distinct constraints which give same marks to same candidates. ⇒ Patterns supported by such equivalent constraints would benefit from multiplicity.

b. *Extensional equivalence.* In particular language, intensionally distinct constraints may be extensionally equivalent, and would thus *act* like copies of each other. E.g., voicing assimilation in French could be described either [voice] assimilation or [Laryngeal] assimilation; a ban on [$+low$] vowels in some context is also a ban on [$+low$, $-high$] vowels. Very frequent in Subtree learner.

c. *Constraint overlap.* It could also happen that the constraint set contains constraints which, though they do not act identically all the time, do so often enough to have the effect of multiplicity. Example: The PCL as described in the Pater UNC colloq talk uses a "tier-sensitive bigram" constraint schema which is engineered to induce multiple overlapping constraints in response to within-tier dependencies, but not between-tier ones.

d. *Inducer timeout.* If there are more ways to express Constraint A than Constraint B, then Constraint A is likely to be found first, and in a time-limited learner B may not be found at all.

## References

Albright, A. and B. Hayes (2002). Modelling English past tense intuitions with minimal generalization. In M. Maxwell (Ed.), *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology, Philadelphia, July 2002*, pp. ??–?? Association for Computational Linguistics.

Bach, E. and R. T. Harms (1972). How do languages get crazy rules? In R. P. Stockwell and R. K. S. Macaulay (Eds.), *Linguistic change and generative theory*, Chapter 1, pp. 1–21. Bloomington: Indiana University Press.

Becker, M., N. Ketrez, and A. Nevins (2007). Where and why to ignore lexical patterns in Turkish obstruent alternations. Handout, 81st annual meeting of the Linguistic Society of America, Anaheim, California.

Boersma, P. (1998). *Functional Phonology: formalizing the interactions between articulatory and perceptual drives.* Ph. D. thesis, University of Amsterdam.

Boersma, P. and B. Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry 32*, 45–86.

Boersma, P. and J. Pater (2007, October). Constructing constraints from language data: the case of Canadian English diphthongs. Handout, NELS 38, University of Ottawa.

Burzio, L. (1999). Surface-to-surface morphology: when your representations turn into constraints. MS, Department of Cognitive Science, Johns Hopkins University. ROA-341.

Carpenter, A. C. (2005). Acquisition of a natural vs. an unnatural stress system. In A. Brugos, M. R. Clark-Cotton, and S. Han (Eds.), *Papers from the 29th Boston University Conference on Language Development (BUCLD 29)*, Somerville, pp. 134–143. Cascadilla Press.

Eiben, A. E. and J. E. Smith (2003). *Introduction to evolutionary computing.* Berlin: Springer.

Gildea, D. and D. Jurafsky (1995). Automatic induction of finite-state transducers for simple phonological rules. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics (ACL-95), Cambridge, Massachusetts*, pp. 9–15. Association for Computational Linguistics.

Gussenhoven, C. and H. Jacobs (2005). *Understanding phonology* (2nd ed.). Understanding Language Series. London: Hodder Arnold.

Hale, M. and C. A. Reiss (2000). 'substance abuse' and 'dysfunctionalism': current trends in phonology. *Linguistic Inquiry 31*(1), 157–169.

Hayes, B. (1999). Phonetically driven phonology: the role of optimality in inductive grounding. In M. Darnell, E. Moravcsik, M. Noonan, F. Newmeyer, and K. Wheatly (Eds.), *Functionalism and Formalism in Linguistics*, Volume 1: General Papers, pp. 243–285. Amsterdam: John Benjamins.

Hayes, B. (2004). Phonological acquisition in Optimality Theory: the early stages. In R. Kager, J. Pater, and W. Zonneveld (Eds.), *Constraints in phonological acquisition*, Chapter 5, pp. 158–203. Cambridge, England: Cambridge University Press.

Hayes, B. and C. Wilson (2008). A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39*(3), 379–440.

Hayes, B., K. Zuraw, P. Siptár, and Z. Londe (2008, September). Natural and unnatural constraints in Hungarian vowel harmony. MS, Department of Linguistics, University of California, Los Angeles.

Heinz, J. (2007). Learning phonotactic grammars from surface forms. In D. Baumer, D. Montero, and M. Scanlon (Eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics*, Somerville, pp. 186–194. Cascadilla.

Hsieh, H. (1976). On the unreality of some phonological rules. *Lingua 38*, 1–19.

Jäger, G. (to appear). Maximum Entropy models and Stochastic Optimality Theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, and A. Zaenen (Eds.), *Architectures, rules, and preferences: a festschrift for Joan Bresnan.* Stanford: CSLI.

Kager, R., N. Boll-Avetisyan, and C. Ao (2008, November). Gradient phonotactic constraints for speech segmentation in a second language. Poster presented at the Boston University Conference on Language Development.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology 25*(1), 83–127.

Pater, J. (2008). Handout from colloquium talk, Department of Linguistics, University of North Carolina, Chapel Hill.

Pater, J., E. Moreton, and M. Becker (2008, November). Simplicity biases in structured statistical learning.

Poster presented at the Boston University Conference on Language Development.

Saffran, J. R. and E. D. Thiessen (2003). Pattern induction by infant language learners. *Developmental Psychology 39*(3), 484–494.

Schane, S. A., B. Tranel, and H. Lane (1974). On the psychological reality of a natural rule of syllable structure. *Cognition 3*(4), 351–358.

Smith, J. L. (2002). *Phonological augmentation in prominent positions.* Ph. D. thesis, University of Massachusetts, Amherst.

Smith, J. L. (2004). Making constraints positional: towards a compositional model of con. *Lingua 114*(2), 1433–1464.

Smith, J. L. (2006). Representational complexity in syllable structure and its consequences for Gen and Con. MS, Department of Linguistics, University of North Carolina, Chapel Hill. ROA-800.

Wilson, C. (2003). Experimental investigation of phonological naturalness. In G. Garding and M. Tsujimura (Eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL 22)*, Somerville, pp. 533–546. Cascadilla Press.

Wilson, C. (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science 30*(5), 945–982.

Zimmer, K. E. (1969). Psychological correlates of some Turkish morpheme structure constraints. *Language 45*(2), 309–321.