

# Mathematical models of variation and change: an application to an English stress shift

Partha Niyogi<sup>1</sup>    Morgan Sonderegger<sup>2</sup>

<sup>1</sup>Departments of Computer Science and Statistics, University of Chicago

<sup>2</sup>Department of Computer Science, University of Chicago

Northwestern University  
December 2, 2008

**Introduction**

**Data: English N/V pairs**

**Sources of change**

**Models**

**Discussion**

## Introduction

- ▶ Given variation  $A/B$  between old form  $A$  and new form  $B$ , can have:
  1. Variation disappears ( $A \sim B \rightarrow A$ )
  2. Variation persists ( $A \sim B \rightarrow A/B$ )
  3. Change occurs ( $A \sim B \rightarrow B$ )
- ▶ Fine phonetic variation extensive, yet (3) uncommon. **How do variation and change coexist?**
- ▶ The actuation problem, restated:
  1. Why does language change occur at all?
  2. Why does it arise from variation?
  3. What determines whether a pattern of variation is stable or unstable (leads to change)?
- ▶ Explore through diachronic dataset + mathematical modeling.

## Summary

- ▶ Data: Observe *coupling* pattern in English N/V pairs: {N=1,V=1}, {N=1,V=2}, {N=2,V=2}, but never {N=2,V=1}
- ▶ Q: How can  $\{2,1\}$  persist diachronically?
- ▶ Claim: Different kinds of plausible learning algorithms for same pattern can give very different diachronic dynamics.
- ▶ Four model classes considered here, based on ling literature:
  1. Mistransmission
  2. Discarding
  3. Regularization
  4. Coupling
- ▶ All offer plausible explanations for N/V pattern, only (4) (mostly) works.

## Previous work

- ▶ Computational studies of language change mushrooming (Baker, 2008)
- ▶ Often (a) simulation-based, (b) few (1–5) models
- ▶ Modeling linguistic populations with intraspeaker variation: Harrison et al. (2002); Yang (2002); Mitchener (2005); Niyogi (2006); Daland et al. (2007); Troutman et al. (2008), ...
- ▶ Advantages to (a, b), but hard to reason exactly about source of model behavior, test predictions.
- ▶ Complementary perspective: consider larger number of simpler models; find source of meaningful patterns from more complicated models.
- ▶ Tradeoff: simplified network, lexicon structure
- ▶ By exploring “landscape” of models, connect model and dataset properties.

## Data: English N/V pairs

- ▶ English 2-syllable noun/verb pairs have variable stress:

	N	V	
{1, 1}	óσ	óσ	(elbow, fracture)
{1, 2}	óσ	σó	(consort, protest)
{2, 2}	σó	σó	(police, review)

- ▶ Variation between dialects (British address, Indian delay)
- ▶ Ongoing variation within dialects: US perfume, research, ally...

## What kind of variation?

- ▶ Linguistic variation can be inter- or intra-speaker.
- ▶ Experiment: find speakers on (mostly) National Public Radio.
- ▶ 33 speakers (15 F/18 M), 48 stories.<sup>1</sup>
- ▶  $\geq 5$  tokens per word per speaker.

Word	1 only	2 only	Var	Spkrs
research (N)	0.53	0.12	0.35	17
perfume (N)	0.22	0.44	0.33	9
address (N)	0.4	0.4	0.2	5

- ▶ Intraspeaker variation, interesting structure.

---

<sup>1</sup><http://people.cs.uchicago.edu/~morgan/diatones/radioStories.pdf>

## Diachronic behavior

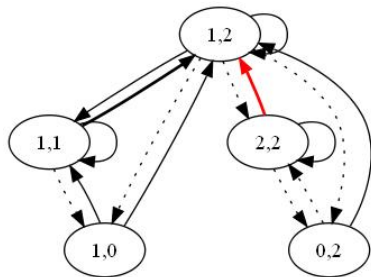
- ▶ In random subset of N/V pairs, most do not change pronunciation 1700-2000:  $\{1, 1\}$ ,  $\{1, 2\}$ ,  $\{2, 2\}$  are *stable states*.
- ▶ Sherman (1975) gives list  $\mathcal{L}$  of 149 which did change, most  $\rightarrow \{1, 2\}$ .
- ▶ What does change in  $\mathcal{L}$  look like?
- ▶ Dataset:  $\mathcal{L}$  pronunciation in 76 dictionaries,<sup>2</sup> 1550-2007 (most 1700–), collected by Sherman (1550–1800), us (1800–2007).
- ▶ Variation often recorded:  
*"Although all the orthoepists accent this word on the second syllable, yet we often hear it pronounced with the accent on the first."* (Worcester, 1859)

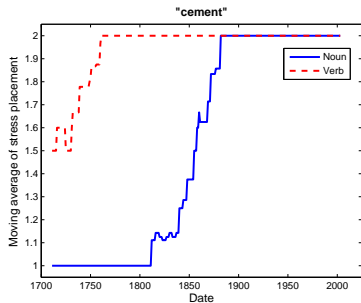
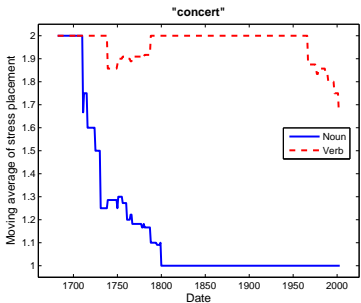
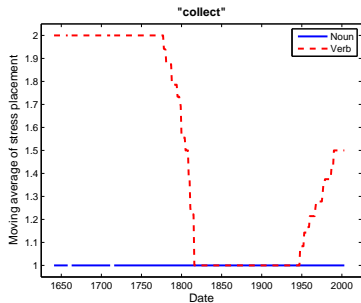
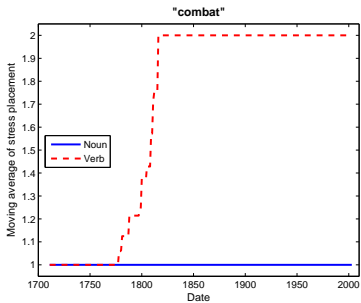
---

<sup>2</sup><http://people.cs.uchicago.edu/~morgan/diatones/dictList.pdf>

## Trajectory dynamics I

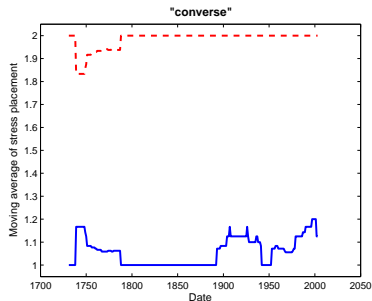
- ▶ Plot moving average of  $N$ ,  $V$  pronunciations for each pair, 50 year window, write as  $p(w, t)$ .
- ▶ “Endpoint” at  $t$  if all prons in window at  $t$  agree ( $p(w, t) = \{1, 1\}, \{1, 2\},$  or  $\{2, 2\}$ ).
- ▶ Changes observed (by freq)
  1.  $\{2, 2\} \rightarrow \{1, 2\}$
  2.  $\{1, 1\} \rightarrow \{1, 2\}$
  3.  $\{1, 2\} \rightarrow \{1, 1\}$
  4.  $\{1, 2\} \rightarrow \{2, 2\}$
- ▶ Often sudden, sigmoidal.



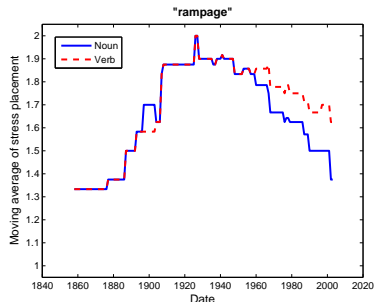


## Trajectory dynamics II

- ▶ Short-term variation common.



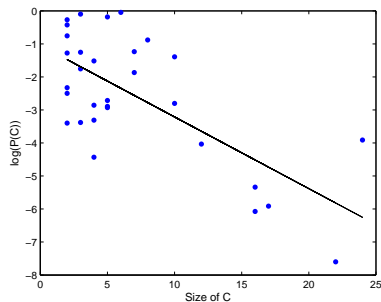
- ▶ Long-term variation in N and V together rare.



- ▶  $\{2, 1\}$  never occurs.
- ▶ Trajectories stay near stable states.

## Measuring analogy

- ▶ How much do N/V pairs with same prefix “change like” each other?
- ▶ Define distance metric  $p(C)$  over subsets  $C \subset \mathcal{L}$ :  
 $p(\text{'re'}) = 0.011$ ,  $p(\text{'com'}) = 0.067$ ..
- ▶ Measures how dissimilar trajectories are for all  $w_1, w_2 \in C$ , normalizes.



## Trajectory dynamics: other

- ▶ Trajectories are largely along 1D subspace  $\{1, 1\} \leftrightarrow \{1, 2\} \leftrightarrow \{2, 2\}$ :<sup>3</sup>

$p(w, t)$	{1,1}	{1,var}	{1,2}	{var,2}	{2,2}	{var,var}
% entries	0.07	0.05	0.57	0.07	0.22	0.02

- ▶ i.e. usually **only one of N or V changes at a time**.
- ▶ Frequency effects observed, not discussed here.

---

<sup>3</sup>“Var” = both forms reported

## Sources of N/V patterns: previous work

- ▶ Generalization over English lexicon that N stress farther left than V stress (e.g. Ross (1973)).
- ▶ Why change  $\rightarrow \{1, 2\}$ ?
- ▶ Kelly (1988a, 1988b, 1989), Davis & Kelly (1997): Perception of  $\sigma\sigma$  N biased  $\rightarrow \acute{\sigma}\sigma$ , V biased  $\rightarrow \sigma\acute{\sigma}$ .
  - ▶ Why? N more common in trochaic-biasing ( $\acute{\sigma}\sigma$ ) than iambic-biasing ( $\sigma\acute{\sigma}$ ) contexts, v.v. for verbs.
  - ▶ “Use the colvane proudly”/ “The proud colvane refused”
- ▶ Guion et al. (2003): Disyllabic nonsense word stress assignment based on lexical class  $>$  syllable structure  $>$  lexical neighborhood.<sup>4</sup>
  - ▶ % Initial stress N  $>$  V

---

<sup>4</sup>CVVCVCC  $>$  CVCVCC  $>$  CVCVC  $>$  CVCVVC

## Proposed factors in change

- ▶ **Mistransmission** (Ohala (1981); Blevins (2004)..): Speaker intends A, hearer perceives B.  
N/V: Kelly studies
- ▶ **Analogy/lexicon** (Historical linguists, Pierrehumbert (2001); Bybee (2002)): Pron of form influenced by other forms  
N/V: Prefix class trajectory similarity, Guion et al., N<V stress generalization
- ▶ **Regularization** (Singleton & Newport (2004); Hudson Kam & Newport (2005)): Learners have categoricity bias.  
N/V: NPR data?
- ▶ **Filtering** (Pearl & Weinberg (2007), J. Morgan): Learners filter input (e.g. for unambiguous data).

## Models: Outline

- ▶ N/V facts to be modeled
- ▶ Summary of all models
- ▶ Dynamical systems introduction
- ▶ Worked models
  - ▶ Mistransmission
  - ▶ Discarding
  - ▶ Regularization
  - ▶ Coupling
- ▶ Other models: overview

## Summary of diachronic facts to be modeled

- ▶ Stable variation in both N and V rare; stable variation in N or V common.
- ▶ Words sharing prefix have more similar trajectories (provided prefix class not small).
- ▶  $\{2,1\}$  never occurs,  $\{1,1\}$ ,  $\{1,2\}$ ,  $\{2,2\}$  do
- ▶ Trajectories move in  $\{1,1\} \leftrightarrow \{1,2\} \leftrightarrow \{2,2\}$  subspace.
- ▶ Change to  $\{1,2\}$  common, from  $\{1,2\}$  not.

## Models summary

- ▶ No coupling (single forms)
  - ▶ **Mistransmission** × **discarding** × input type × Regularization
  - ▶ Input type: fixed or Poisson
  - ▶ Regularization: thresholding, **frequency boosting**
  - ▶ Bayesian learners: MAP, posterior mean
- ▶ Coupling between N and V forms
  - ▶ By grammar: learn probs for  $\{1, 1\}$ ,  $\{1, 2\}$ ,  $\{2, 1\}$ ,  $\{2, 2\}$
  - ▶ **By constraint**:  $P(N=2) < P(V=2)$
- ▶ Coupling between N/V pairs
  - ▶ Prior/data competition
- ▶ Today, selection.

## Discrete dynamical systems

- ▶ Common in population biology. Niyogi & Berwick (1995); Niyogi (2006) → linguistic populations
- ▶ System state  $\alpha$  evolves by  $\alpha_{t+1} = f(\alpha_t)$ .
- ▶ Dynamical systems viewpoint: examine limiting ( $t \rightarrow \infty$ ) behavior.
- ▶ Find *fixed points* where  $f(\alpha^*) = \alpha^*$ . Can be
  - ▶ Stable: For  $\alpha_0$  near  $\alpha^*$ ,  $\alpha_t \rightarrow \alpha^*$ . ( $|f'(\alpha^*)| < 1$ )
  - ▶ Unstable: For  $\alpha_0$  near  $\alpha^*$ ,  $\alpha_t \not\rightarrow \alpha^*$ . ( $|f'(\alpha^*)| > 1$ )
- ▶ *Bifurcation*: number or stability of FPs of  $f$  changes as system parameters which determine  $f$  varied.
- ▶ Goal: For given  $f$ , find FPs and stabilities. Given system parameters determining  $f$ , find bifurcations.
- ▶ Worked examples below..

## Observed N/V behavior as DS desired properties

- ▶ Sudden change: **Bifurcations**
- ▶ Different dialects, different forms:<sup>5</sup> **Multistability** (multiple stable fixed points at once)
- ▶  $\{1,1\}$ ,  $\{1,2\}$ ,  $\{2,2\}$ : **Stable states** (for some system parameter values)
- ▶  $\{2,1\}$ : **Unstable state**
- ▶  $\{1,1\} \leftrightarrow \{1,2\} \leftrightarrow \{2,2\}$  trajectories:  $\alpha_t$  trajectories **lie in 1D subspace.**

---

<sup>5</sup>In this interpretation

## Model assumptions

- ▶ Simplifying assumptions:
  - ▶ *Discrete generations*: Learners in generation  $n$  learn from generation  $n - 1$ .
  - ▶ *Full connectivity*: Each example a learner in generation  $n$  hears is equally likely to come from any member of generation  $n - 1$ .
  - ▶ *Infinite populations*: Each generation has infinitely many members.
  - ▶ *Fixed input*: Each learner receives constant  $N$  examples.
- ▶ None sacred, see (Niyogi, 2006) for variations.
- ▶ Gradience: Probabilities  $\alpha \in [0, 1]$  over forms learned.

## Model 1: Probability matching with mistransmission

- ▶ Single form, each speaker keeps  $\hat{\alpha} \in [0, 1]$  = probability of producing form 2 vs form 1.
- ▶  $\alpha_t$ : probability random example from generation  $t$  produced as form 2.
- ▶ Define mistransmission probabilities:

$$a = P(\text{Hear 1} \mid 2 \text{ intended}), \quad b = P(\text{Hear 2} \mid 1 \text{ int})$$

- ▶ Probability of learner at  $t + 1$  hearing 2 example now

$$p_2(t) = \alpha_t(1 - a) + (1 - \alpha_t)b$$

- ▶ Algorithm:
  1. Learner in generation  $t$  receives data  $X_1, \dots, X_N$  of which  $k$  are form 2,  $N - k$  form 1.
  2. Probability matches:  $\hat{\alpha} = \frac{k}{N}$ .

- ▶  $\hat{\alpha}_t$ : random variable corresponding to learners in generation  $t + 1$ :  $\alpha_{t+1} = E(\hat{\alpha}_t)$ .
- ▶  $k \sim \text{Binom}(\alpha_t, N)$ , so

$$P(\hat{\alpha}_t = \frac{k}{N}) = P(k) = \binom{N}{k} p_2(t)^k (1 - p_2(t))^{N-k}$$

- ▶ Averaging over generation  $t$  learners gives:

$$\begin{aligned} \alpha_{t+1} = E(\hat{\alpha}_t) &= \sum_{k=0}^N P(\hat{\alpha} = \frac{k}{N}) \cdot \frac{k}{N} \\ &= \sum \binom{N}{k} p_2(t)^k (1 - p_2(t))^{N-k} \frac{k}{N} \\ &= \frac{1}{N} (p_2(t)N) = p_2(t) \\ \implies \alpha_{t+1} &= \alpha_t(1 - a) + (1 - \alpha_t)b \end{aligned}$$

- ▶ So  $f(x) = x(1 - a) + (1 - x)b$ .
- ▶ Solve  $f(x) = x \implies$  unique fixed point if  $a + b > 0$ :

$$x^* = \frac{b}{a + b}$$

- ▶  $f'(x^*) = 1 - a - b \implies$  stable (for  $a + b > 0$ ).
- ▶  $a + b = 0$ :  $\alpha_{t+1} = \alpha_t$ , identity map.

## Model 2: Discarding

- ▶ One interpretation of “filtering”: no mistransmission, but each example can be heard as 1, 2, or discarded.
- ▶ Discarding probabilities:

$$r_i = P(\text{discarded} \mid i \text{ intended}) \quad (i = 1, 2)$$

- ▶ Now

$$P_1(t) = (1 - \alpha_{t-1})(1 - r_1), \quad P_2 = \alpha_{t-1}(1 - r_2)$$

- ▶ Algorithm: Learner at  $t$  gets  $k_1, k_2$  1 and 2 examples, sets

$$\hat{\alpha} = \begin{cases} \frac{K_2}{K_1 + K_2} & \text{if } K_1 + K_2 > 0 \\ \frac{1}{2} & \text{if } K_1 + K_2 = 0 \end{cases}$$

- ▶  $(k_1, k_2) \sim \text{Multinom}(p_1(t), p_2(t), N)$  across learners at  $t$ .

$$\alpha_t = E(\hat{\alpha}_t) = P(k_1 + k_2 = 0) \cdot \frac{1}{2} + \sum_{k_1+k_2>0}^N \underbrace{P(k_1, k_2)}_{\text{Multinom}} \frac{k_2}{k_1 + k_2}$$

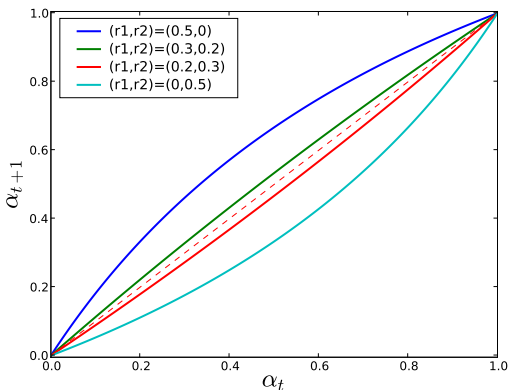
- ▶ For large  $N$ , get

$$\alpha_{t+1} = \frac{\alpha_t(1 - r_2)}{(1 - r_1) + \alpha_t(r_1 - r_2)}$$

- ▶ Solve  $\alpha_{t+1} = \alpha_t$ , FPs at 0,1:

$$\alpha_+ = 1 \text{ stable for } r_1 > r_2, \quad \alpha_- = 0 \text{ stable for } r_1 < r_2$$

$\alpha_{t+1} = \frac{\alpha_t(1-r_2)}{(1-r_1)+\alpha_t(r_1-r_2)}$ : Slopes at 0, 1 determine stability.



- ▶ **Bifurcation** at  $r_1 = r_2$ : FPs switch stability.
- ▶ For  $N$  smaller, same behavior with  $x_+$ ,  $x_-$  shifted off 0,1.

## Mistransmission, Discarding: summary

- ▶ Mistransmission shifts fixed points, but doesn't cause bifurcations
- ▶ Discarding changes fixed point stabilities (bifurcation), doesn't shift them.
- ▶ Basically true for all models considered in which mistransmission or discarding introduced.

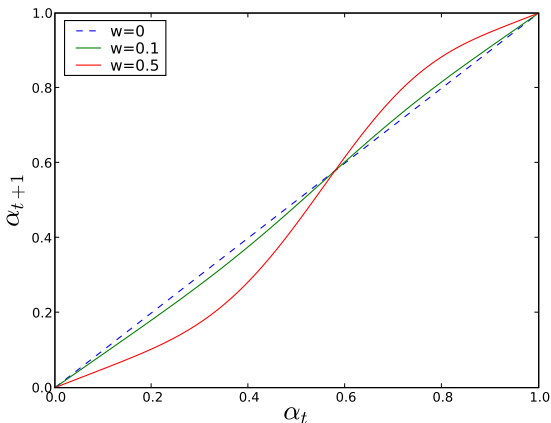
## Model 3: Regularization I: frequency boosting

- ▶ “Frequency boosting” (Singleton & Newport, 2004): Learners estimate closer to 0, 1 than probability match  $\frac{k}{N}$ .
- ▶ Implement as: Weight  $w \in [0, 0.5)$ .
- ▶ Algorithm:
  1. Learner gets  $k$  form 2,  $N - k$  form 1 examples.
  2. Sets  $\hat{\alpha} = \begin{cases} \frac{k}{N}(1 - w) & : \frac{k}{N} \leq 0.5 \\ \frac{k}{N}(1 - w) + w & : \frac{k}{N} > 0.5 \end{cases}$
  3. Weighted average of  $\frac{k}{N}$  and nearest endpoint.

► Dynamics:

$$\begin{aligned}\alpha_{t+1} &= E[\hat{\alpha}_t] = \sum_{k \leq \frac{N}{2}} P(k) \frac{k}{N} (1-w) + \sum_{k > \frac{N}{2}} P(k) \left[ \frac{k}{N} (1-w) + w \right] \\ &= \sum_{k \leq \frac{N}{2}} \binom{N}{k} \alpha_t^k (1-\alpha_t)^{N-k} \frac{k}{N} (1-w) \\ &\quad + \sum_{k > \frac{N}{2}} \binom{N}{k} \alpha_t^k (1-\alpha_t)^{N-k} \left[ \frac{k}{N} (1-w) + w \right] \\ \alpha_{t+1} &= \alpha_t + w \sum_{k > \frac{N}{2}} \binom{N}{k} \alpha_t^k (1-\alpha_t)^{N-k}\end{aligned}$$

$$\alpha_{t+1} = \alpha_t + w \sum_{k > \frac{N}{2}} \binom{N}{k} \alpha_t^k (1 - \alpha_t)^{N-k}$$



- ▶ For  $w > 0$ , fixed points are 0, 1 (stable),  $\alpha^*$  (unstable):  
**Bistability**

## Model 3.1: Regularization II

- ▶ Regularization model II (thresholding):

$$\epsilon > 0, \text{ set } \hat{\alpha} = \begin{cases} 0 & : \frac{k}{N} < \epsilon \\ 1 & : \frac{k}{N} > 1 - \epsilon \\ \frac{k}{N} & : \text{otherwise} \end{cases}$$

- ▶ Same dynamics as frequency boosting.

## Regularization: Summary

- ▶ Get bistability for arbitrarily small bias: need only  $\epsilon, w > 0$ .
- ▶ Convergence of model dynamics

## Coupling models: motivation

- ▶ One-form case: probability matching possible.
- ▶ Coupled N and V forms: Not possible – learner sees individual forms, not  $\{N,V\}$  pairs.
- ▶ Given evidence for  $\{N,V\}$  generalizations, inference necessary.
- ▶ To explain: Why can  $\{1,1\}$ ,  $\{1,2\}$ ,  $\{2,2\}$  occur, but not  $\{2,1\}$ ?

## Model 4: Coupling by constraint

- ▶ Single N/V pair.
- ▶ Let  $\hat{\alpha}$ ,  $\hat{\beta}$  be learned probabilities of producing N, V forms as 2 (final stress).
- ▶  $\alpha_t, \beta_t$  mean of  $\hat{\alpha}, \hat{\beta}$  over generation  $t$ .
- ▶ Constraint:  $\hat{\alpha} < \hat{\beta}$  (c.f. English stress generalization).
- ▶ Algorithm:
  1. Hear  $N_1$  N,  $k_1$  as form 2,  $N_2$  V,  $k_2$  as form 2.
  2. If  $\frac{k_1}{N_1} < \frac{k_2}{N_2}$ , set  $\hat{\alpha} = \frac{k_1}{N_1}, \hat{\beta} = \frac{k_2}{N_2}$ .
  3. Otherwise, set  $\hat{\alpha} = \hat{\beta} = \frac{1}{2} \left( \frac{k_1}{N_1} + \frac{k_2}{N_2} \right)$
- ▶ 2. satisfies optimization problem

$$\text{minimize } \left[ \left( \alpha - \frac{k_1}{N_1} \right)^2 + \left( \beta - \frac{k_2}{N_2} \right)^2 \right] \text{ s.t. } \alpha \leq \beta$$

- ▶  $k_1 \sim \text{Binom}(\alpha_t, N_1)$ ,  $k_2 \sim \text{Binom}(\alpha_t, N_2)$  (independent), so

$$P(k_1, k_2) = \binom{N_1}{k_1} \binom{N_2}{k_2} \alpha^{k_1} (1 - \alpha)^{N_1 - k_1} \beta^{k_2} (1 - \beta)^{N_2 - k_2}$$



$$\begin{aligned} \alpha_{t+1} = E[\hat{\alpha}_t] &= \sum_{\frac{k_1}{N_1} < \frac{k_2}{N_2}} P(k_1, k_2) \frac{k_1}{N_1} \\ &+ \sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P(k_1, k_2) \frac{1}{2} \left( \frac{k_1}{N_1} + \frac{k_2}{N_2} \right) \\ \dots &= \alpha_t - \underbrace{\frac{1}{2} \sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P(k_1, k_2) \left( \frac{k_1}{N_1} - \frac{k_2}{N_2} \right)}_A. \end{aligned}$$

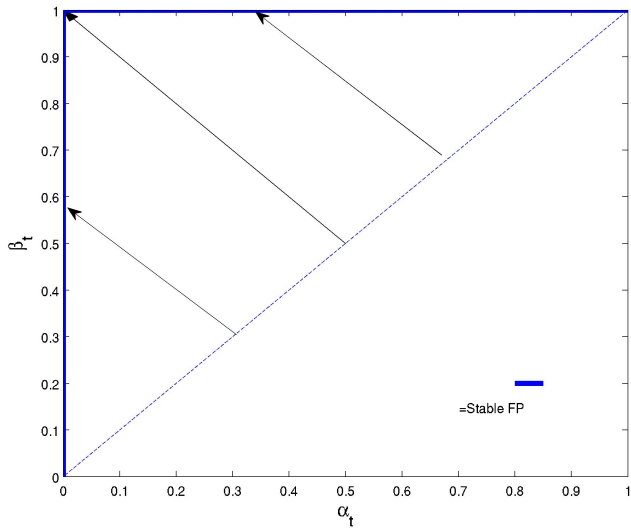
- ▶ End up getting

$$\alpha_{t+1} = \alpha_t + \frac{A}{2}, \quad \beta_{t+1} = \beta_t - \frac{A}{2}$$

- ▶  $\alpha_{t+1} + \beta_{t+1} = \alpha_t + \beta_t$ :  
Trajectories along constant  $\alpha + \beta$  lines
- ▶ Fixed points: all  $(0, x)$ ,  $(x, 1)$ ,  $x \in [0, 1]$ .<sup>6</sup>
- ▶ i.e.  $\{N=1, V=\text{var}\}$ ,  $\{N=\text{var}, V=2\}$

---

<sup>6</sup>Notation:  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 1)$  are the fixed points corresponding to  $\{1, 1\}$ ,  $\{1, 2\}$ ,  $\{2, 2\} \equiv \{N=1, V=1\}$ ,  $\{N=1, V=2\}$ ,  $\{N=2, V=2\}$



## Model 5: Prior/data competition

- ▶  $k_1, k_2, N_1, N_2$  as above, but now want to estimate probabilities of using each N/V *grammar*: could set

$$\hat{P}_{11} = \frac{k_1}{N_1} \frac{k_2}{N_2}, \quad \hat{P}_{12} = \frac{k_1}{N_1} \frac{N_2 - k_2}{N_2}, \quad \hat{P}_{22} = \frac{N_1 - k_1}{N_1} \frac{N_2 - k_2}{N_2}$$

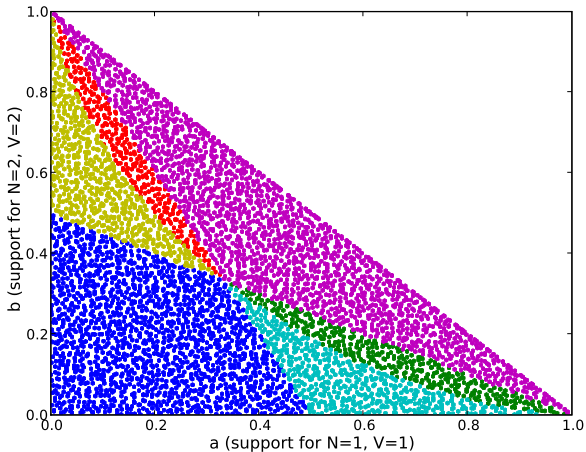
- ▶ Say  $\exists$  “prior” probabilities for grammars:  $a_{11}, a_{12}, a_{22}, a_{21}$ .
- ▶ E.g. % N/V pairs with same prefix following pattern  $ij$ .
- ▶ To produce N or V, choose a grammar according to  $a_{ij}$  and a grammar according to  $\hat{P}_{ij}$ , repeat until predictions agree.
- ▶ For given learner, gives *effective*  $\hat{\alpha}$  and  $\hat{\beta}$  (for  $P(N = 2), P(V = 2)$ ).
- ▶ Dynamical system in  $(\alpha_t, \beta_t)$  as above.

- ▶ Skipping details...
- ▶ Take  $a_{21}$  small ( $\{2,1\}$  not in lexicon).
- ▶ Define constants

$$B = \left( \frac{a_{22}N_1}{a_{22} - a_{12} + a_{12}N_1} \right), \quad A = \left( \frac{a_{11}N_2}{a_{11} - a_{12} + a_{12}N_2} \right)$$

- ▶ Six FP regions:
  1.  $a_{11}, a_{22} < a_{12}$ :  $(0, 1)$  stable
  2.  $a_{22} > a_{12}, AB < 1$ :  $(0, 1), (1, 1)$  stable
  3.  $a_{11} < a_{12} < a_{22}, AB > 1$ :  $(1, 1)$  stable
  4.  $a_{11}, a_{22} > a_{12}$ :  $(0, 0), (1, 1)$  stable
  5.  $a_{22} < a_{12} < a_{11}, AB > 1$ :  $(0, 0)$  stable
  6.  $a_{11} > a_{12}, AB < 1$ :  $(0, 0), (0, 1)$  stable
- ▶ Bifurcations depend on  $N_1, N_2, a_{ij}$

- Phase diagram in  $(a_{11}, a_{22})$  with  $N_1 = 5, N_2 = 10$ :



- Stable FPs:
 

$(0,1)$	$(0,1), (1,1)$	$(0,0)$
$(0,0), (1,1)$	$(1,1)$	$(0,0), (0,1)$

## Coupling models: summary

- ▶ N stress  $<$  V stress constraint allows too many stable states
- ▶ Also, wrong trajectories.
- ▶ Prior/data competition: Closest to observed behavior
- ▶ Bifurcations where  $\{1,1\}$  or  $\{0,0\}$  suddenly become unstable  $\implies$  change to  $\{0,1\}$ .

## Other models: Highlights

- ▶ **Poisson input:** Instead of fixed number  $N$  examples, learners get  $N \sim \text{Poisson}(\lambda)$ 
  - ▶ Large  $N$ : no difference
  - ▶ Small  $N$ : Dynamics determined by  $N = 0$  behavior.
- ▶ **Bayesian learners:** (No coupling) Learners start with  $\text{Beta}(A, B)$  prior for  $\alpha \implies$  posterior  $P(\alpha | k, N)$ . Then must estimate  $\hat{\alpha}$ :
  - ▶  $\hat{\alpha} =$  Posterior mode (MAP): Bifurcations in  $A, B$ .
  - ▶  $\hat{\alpha} =$  Posterior mean:<sup>7</sup>Single fixed point
- ▶ **Coupling by grammar:** Store  $\{N, V\}$  probabilities, not  $N$  and  $V$  probabilities (c.f. Yang (2002))
  - ▶ Single fixed point

---

<sup>7</sup>Equivalent to sampling from the posterior

## Model properties summary

	Bifurcs	Mult FPs	*{2,1}
Mistransmission	x	x	x
Discarding	✓	x	x
Regularization	✓/x	✓	x
Coupling (constraint)	x	✓	✓
Coupling (prior/data)	✓	✓	✓
Coupling (by grammar)	x	x	x/✓

- ▶ Mistransmission moves FPs around, can determine *direction* of dynamics, but doesn't itself give interesting dynamics.
- ▶ Basically need to rule out {2,1} to have it not occur, but *how* coupling done matters for rest of dynamics.

## Discussion

- ▶ What differentiates models that “work”  $\equiv$  show several properties seen in natural variation and change?
- ▶ Proposal: **Need bias in learning algorithm, not (just) bias in data.**
- ▶ Practical implications:
  1. Mistransmission-type models (misperception, coarticulation, overcompensation..) are the most commonly proposed sources of sound change... but here, don't (alone) give right dynamics.
  2. To get right N/V behavior, coupling N and V dynamics necessary: Suggests patterns of change can reveal underlying learner biases, c.f. recent work (Griffiths, Kirby, Newport..)

- ▶ Feasible and worthwhile to build mathematical models, dataset, go back and forth.
- ▶ Future directions:
  1. Social network structure, finite populations
  2. Structured lexicon (as graph)
  3. “Rational models” for coupling: Learner assumes PDF on forms, makes Bayesian or maximum likelihood (etc.) estimate.

## References

- Baker, A. (2008) Computational Approaches to the Study of Language Change. *Language and Linguistics Compass*, 2:289–307.
- Blevins, J. (2004) *Evolutionary Phonology* Cambridge University Press.
- Bybee, J. (2002) Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14: 261-290.
- Daland, R., Sims, A.D., and Pierrehumbert, J.B. (2007). Much ado about nothing: A social network model of Russian paradigmatic gaps. *Proceedings of 47th Annual Meeting of the Association for Computational Linguistics*.
- Davis, S.M. and Kelly, M.H. (1997) Knowledge of the English Noun–Verb Stress Difference by Native and Nonnative Speakers. *Journal of Memory and Language*, 36:445–460.
- Guion, SG and Clark, JJ and Harada, T. and Wayland, RP (2003). Factors Affecting Stress Placement for English Nonwords include Syllabic Structure, Lexical Class, and Stress Patterns of Phonologically Similar Words *Language and Speech*, 46: 403–427.
- K. Harrison, D., Dras, M. and Kapicioglu, B. Agent-Based Modeling of the Evolution of Vowel Harmony. In *Proceedings of NELS 32*, ed. M. Hirotani.
- Hudson Kam, C.L. and Newport, E.L. (2005) Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development*, 1:151–195.
- Kelly, M. (1988) Phonological biases in grammatical category shifts. *Journal of Memory and Language*, 27: 343-358.
- Kelly, M. (1988). Rhythmic alternation and lexical stress differences in English. *Cognition*, 30:107–137.
- Kelly, M. (1989) Rhythm and language change in English. *Journal of Memory and Language*, 28, 690-710.
- Kelly, M. & Bock, J. (1988). Stress in time. *Journal of Experimental Psychology: Human Performance*, 14, 389-403.
- Mitchener, W. G. (2005) A Simulation of Language Change in the Presence of Non-Idealized Syntax. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition, ACL-2005*.
- Niyogi, P. & Berwick, R. (1995) The logical problem of language change. AI Memo-1516, MIT.
- Niyogi, P. (2006) *The Computational Nature of Language Learning and Evolution*. Cambridge: MIT Press.
- Ohala, J.J. (1981). The listener as a source of sound change. In C.S. Masek, R.A. Hendrick, & M.F. Miller (eds.), *Papers from the Parasession on Language and Behavior*. Chicago: Chicago Ling. Soc. 178–203.
- Pearl, L. and Weinberg, A. (2007) Input Filtering in Syntactic Acquisition: Answers From Language Change Modeling. *Language Learning and Development*, 3:43–72.
- Pierrehumbert, J. (2001) Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (eds.), *Frequency effects and the emergence of lexical structure*. John Benjamins, Amsterdam. 137-157.
- Ross, J.R. (1973). Leftward, ho!. In *Festschrift for Morris Halle*, ed. S.R. Anderson and P. Kiparsky. 166-173.
- Sherman, D. (1975) Noun-verb stress alternation: An example of the lexical diffusion of sound change in English. *Linguistics*, 159: 43-71.
- Singleton, J.L. and Newport, E.L. (2004) When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, 49:370–407.
- Troutman, C., Clark, B., & Goldrick, M. (2008). Social networks and intraspeaker variation during periods of language change. *Penn Working Papers in Linguistics*, 14.1: 325–338.
- Worcester, J. (1859) *A dictionary of the English language*. London; Boston, MA.
- Yang, C. (2002) *Knowledge and Learning in Natural Language*. Oxford University Press.