

# Combining data and mathematical models to study change: An application to an English stress shift

Morgan Sonderegger   Partha Niyogi

University of Chicago

LSA 2010

1/8/2010

## Introduction

- ▶ Three approaches to understanding interplay between variation and change in a speech community/linguistic population:
  1. **Theories of causes** of V&C
  2. **Observed dynamics** of V&C
  3. **Computational framework** to test which theories result in which dynamics.

## Introduction

- ▶ Three approaches to understanding interplay between variation and change in a speech community/linguistic population:
  1. **Theories of causes** of V&C
  2. **Observed dynamics** of V&C
  3. **Computational framework** to test which theories result in which dynamics.
- ▶ Linguists: Mostly (1), (2)
- ▶ Computational linguists, CS, cognitive scientists: Mostly (3)

All three important:

- ▶ **Theories of causes of change**: Motivate computational models.
- ▶ **Data from change**: Test models, make sure not “doomed to success”.
- ▶ **Computational modeling**: Reason about relation between proposed causes (in individuals) and population-level dynamics.

**Larger project** [Sonderegger & Niyogi 2010]:

Relate dynamics of a detailed dataset (2) to a range of mathematical, population-level models (3), inspired by linguistic literature (1).

**Today**: Subset of data and models.

All three important:

- ▶ **Theories of causes of change**: Motivate computational models.
- ▶ **Data from change**: Test models, make sure not “doomed to success”.
- ▶ **Computational modeling**: Reason about relation between proposed causes (in individuals) and population-level dynamics.

**Larger project** [Sonderegger & Niyogi 2010]:

Relate dynamics of a detailed dataset (2) to a range of mathematical, population-level models (3), inspired by linguistic literature (1).

**Today**: Subset of data and models.

- ▶ **Formal framework**: Dynamical systems models of linguistic populations [Niyogi & Berwick 1995, Niyogi 2006]

## Relation to previous work

- ▶ Significant interest in past  $\approx 15$  years in computational models of change [reviews: Niyogi 2006, Baker 2008]
- ▶ Recent interest in combining computational modeling with “real-world” data [Choudhury 2007, Daland et al. 2007, Pearl & Weinberg 2007, Landsbergen 2009].
- ▶ All previous work considers  $< 5$  (usually 1–2) models, sometimes very complex (e.g. agent-based simulation).
- ▶ Our modeling philosophy is different/complementary:
  - ▶ Consider a “landscape” ( $> 10$ ) of relatively simple models, to find source of meaningful patterns, connect model/dataset properties.
  - ▶ Tradeoff: simplified network, lexicon structure

# Summary

## 1. Data

- ▶ Description
- ▶ Dynamics

## 2. Proposed causes of change

- ▶ Mistransmission
- ▶ Analogy

## 3. Models

- ▶ Model 1: Mistransmission
- ▶ Model 2: Analogy
- ▶ Model 3: Mistransmission+Analogy

## 4. Discussion

Introduction

**Data**

Causes

Models

Discussion

## Data: English disyllabic N/V pairs

- ▶ Data: Disyllabic N/V pairs
- ▶ Variable stress:

	N	V	
{1,1}	óσ	óσ	<i>anchor, fracture</i>
{1,2}	óσ	σó	<i>consort, contest</i>
{2,2}	σó	σó	<i>police, review</i>

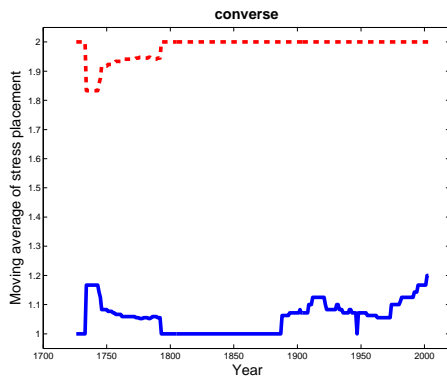
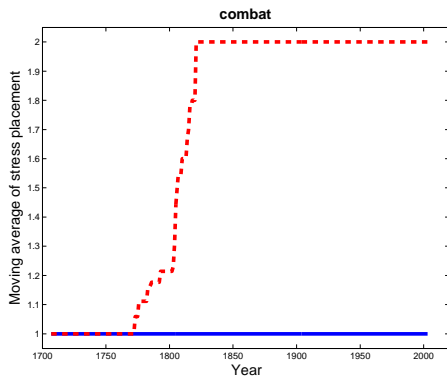
- ▶ **Never {2,1}**
  
- ▶ Sherman (1975):
  - ▶ Considered N/V pairs, ≈1600–1800.
  - ▶ Diachronic stress shift for many pairs, usually to {1,2} (“diatone”).
  
- ▶ Ongoing V&C: *research, perfume, address...*

## Diachronic data

- ▶  $\mathcal{L}$ : Sherman's list of 149 N/V pairs which show V&C.
- ▶ **Our dataset**:  $\mathcal{L}$  stresses reported in 62 historical British dictionaries.
  - ▶ Data collection: Sherman (1550–1800), MS (1800–present).
- ▶ Recorded N, V stress: 1, 2, 1/2, 2/1, NA.  
(1/2=both reported, 1 listed first.)
- ▶  $149 \times 62 \times 2 = 18.5\text{k measurements}$ .
- ▶ Allows detailed description of change.

## Stress trajectories

To visualize V&C, plot moving average of N (blue), V (red) stress:

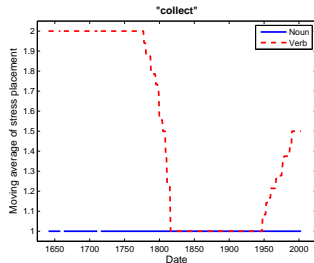
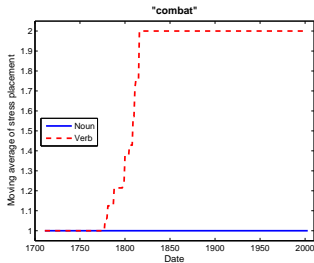


# Observed changes

## Common

$\{2,2\} \rightarrow \{1,2\}$

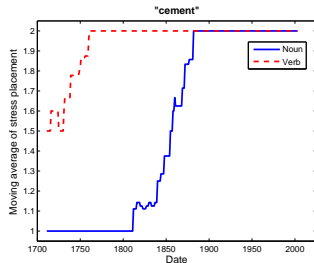
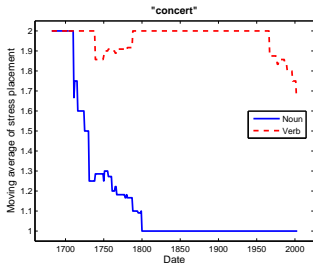
$\{1,1\} \rightarrow \{1,2\}$



## Rarer

$\{1,2\} \rightarrow \{1,1\}$

$\{1,2\} \rightarrow \{2,2\}$



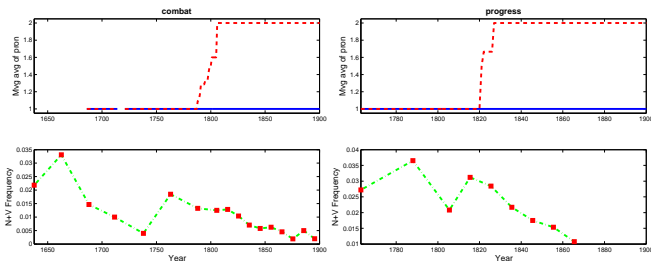
No change between  $\{1,1\}$  and  $\{2,2\}$

## Observed dynamics

- ▶ Also, long-term *stability* at  $\{1,1\}$ ,  $\{1,2\}$ ,  $\{2,2\}$ .
- ▶ Change often sigmoidal and quick, following long-term stability. [c.f. Lightfoot 1991]
- ▶  $\{2,1\}$  never occurs.
- ▶ **Intraspeaker variation**
  - ▶ Dictionaries
  - ▶ Radio corpus (not shown)

# Frequency effects

- ▶ Pairs affected by  $\{2,2\} \rightarrow \{1,2\}$  change have **lower frequency**. [Phillips 1984, Sonderegger 2010/to appear]
- ▶ Some evidence that **falling frequency** (N+V) triggers change:



Change to  $\{1,2\}$  is frequency-dependent

## Trajectory dynamics: Summary

- ▶ Multidirectional, asymmetric change
- ▶ Unobserved changes
- ▶  $\{2,1\}$
- ▶ Frequency dependence

Introduction

Data

**Causes**

Models

Discussion

## Proposed sources of change

- ▶ **Mistransmission** [Ohala 1981 *et seq*, Mowrey & Pagliuca 1995...]
  - ▶ Speaker intends A, hearer perceives B.
- ▶ **Analogy/lexicon** [Historical linguists, Pierrehumbert 2001, Bybee 2002..]
  - ▶ Pron of form influenced by other forms.
- ▶ **Filtering** [Morgan 1986, Pearl 2007]
  - ▶ Learners filter input (e.g. for unambiguous data).
- ▶ **Regularization** [Singleton & Newport 2004, Hudson-Kam & Newport 2005, 2009]
  - ▶ Learners have categoricity bias.

Considered today for N/V case

## Sources of change: Mistransmission

Why is change mostly  $\rightarrow \{1, 2\}$ , and why  $*\{2, 1\}$ ?

1. Productive generalization over English lexicon that N stress farther left than V stress [e.g. Ross 1973].
2. **Biased perception of disyllable stress** [Kelly 1988 *et seq*]
  - ▶ N biased  $\rightarrow \acute{\sigma}\sigma$ , V biased  $\rightarrow \sigma\acute{\sigma}$ .
3. **Biased production** of nonsense disyll stress [Guion et al. 2003]
  - ▶  $N < V$  (% final stress)

## Sources of change: Analogy

- ▶ Prefixed N/V pairs: *contract*, *defect*
- ▶ Not prefixed: *cement*, *police*

Morphological prefix strongly related to change to {1,2}:

1. Almost all pairs in  $\mathcal{L}$  are prefixed. ( $\mathcal{L}$ =pairs which show V/C)
2. Over a *random* list of N/V pairs:
  - ▶ {1,2} N/V pairs: 90% prefixed
  - ▶ All N/V pairs:  $\approx$  38% prefixed
3. N/V pairs sharing a prefix have more similar trajectories than those not sharing a prefix. (Not shown)

Trajectories for different N/V pairs are not independent

## An aside

- ▶ In previous work, N/V stress shift treated as lexical diffusion to {1,2} [Sherman 1975, Phillips 1984]
- ▶ However,
  1. Prefix class effects
  2. Change *from* {1,2}

⇒ **more complicated** than pure LD.
- ▶ More work needed.

Introduction

Data

Causes

**Models**

Discussion

## Dynamical systems: Intro

- ▶ Mathematical framework for systems evolving on a clock, **evolutionary dynamics of populations**.
- ▶ Ported to **linguistic populations** by Niyogi & Berwick [Niyogi & Berwick 1995, Niyogi 2006]
- ▶ **Applications**: Komarova et al. 2001, Yang 2002, Mitchener 2005, Pearl 2007...
- ▶ System state  $\alpha_t$  evolves by  $\alpha_{t+1} = f(\alpha_t)$ :

$$\alpha_0 \xrightarrow{f} \alpha_1 \xrightarrow{f} \alpha_2 \xrightarrow{f} \dots$$

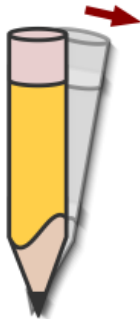
- ▶ **Dynamical systems viewpoint**: Examine limiting ( $t \rightarrow \infty$ ) behavior.

## Fixed points

- ▶ When  $f(\alpha^*) = \alpha^*$ , system is **fixed** at FP  $\alpha_*$ .
- ▶ As  $t$  increases,  $\alpha_t \rightarrow$  a FP.
- ▶ FPs are *stable* or *unstable* under small perturbations:



Stable



Unstable

## Bifurcations

- ▶ Change in the **number or stability of FPs** as system parameter passes a critical value. (a.k.a. phase transition)
- ▶  $\Rightarrow$  Qualitative change in system behavior.
- ▶ Change from  $\alpha_*$ : Bifurcation where FP  $\alpha_*$  **loses stability**.

## Goal of DS analysis

- ▶ Given  $f$ , find FPs and stabilities.
- ▶ Given system parameters determining  $f$ , find bifurcations.

**Bifurcation structure  $\Rightarrow$  possible/impossible changes**

## N/V trajectory dynamics as DS desired properties

1. Sudden change  $\leftrightarrow$  Bifurcations
2.  $\{1,1\}$ ,  $\{1,2\}$ ,  $\{2,2\}$   $\leftrightarrow$  Stable states
  - ▶ For some system parameter values.
3.  $\ast\{2,1\}$   $\leftrightarrow$  Unstable state
4. Observed changes  $\leftrightarrow$  Bifurcation structure
5. Frequency dependence  $\leftrightarrow$  Bifurcation in frequency
  - ▶ Loss of FP  $\{1,2\}$  stability as  $N$  decreased

Goal: Find which models of learning by individuals result in these population-level properties.

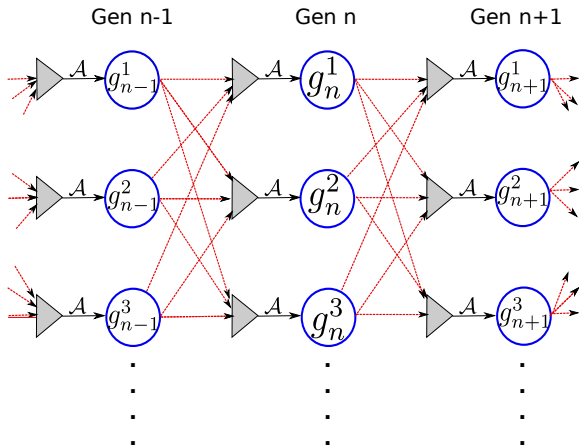
## Model assumptions

- ▶ **Intraspeaker variation**: Learn probabilities of initial vs. final stress for N, V (for a given pair).
- ▶ Simplifying assumptions:
  1. Learners in generation  $n$  learn from generation  $n - 1$ .
  2. Each learner receives same # of examples
  3. Each example equally likely to come from any member of generation  $n - 1$ .
  4. **Each generation has infinitely many members.**

## Model assumptions

- ▶ **Intraspeaker variation**: Learn probabilities of initial vs. final stress for N, V (for a given pair).
- ▶ Simplifying assumptions:
  1. Learners in generation  $n$  learn from generation  $n - 1$ .
  2. Each learner receives same # of examples
  3. Each example equally likely to come from any member of generation  $n - 1$ .
  4. **Each generation has infinitely many members.**
- ▶ Last assumption crucial to dynamics, necessitated by existence of variation in speech communities.
- ▶ Differs from “iterated learning” experiments/simulations, where each generation has **1** member.  
[e.g. Kirby 2000, Kirby et al. 2007, Griffiths & Kalish 2007]

## Models MO



Each learner in Gen  $n$ :

1. Receives data from Gen  $n-1$ .
2. Applies learning algorithm  $\mathcal{A}$  to data.
3. Produces data for Gen  $n+1$ .

## Model notation

- ▶ Each learner receives:

	Total	Heard as $\sigma\acute{\sigma}$	Heard as $\acute{\sigma}\sigma$
N examples	$N_1$	$k_1$	$N_1 - k_1$
V examples	$N_2$	$k_2$	$N_2 - k_2$

- ▶ Applies  $\mathcal{A}$  to learn:

$\hat{\alpha}$  : Prob of producing N as  $\sigma\acute{\sigma}$   
 $\hat{\beta}$  :                   "           V       "

- ▶  $\alpha_t, \beta_t$ :

Probability random N, V example from generation  $t$  produced as  $\sigma\acute{\sigma}$ . (Average over  $\hat{\alpha}$  for learners in Gen  $t$ .)

$\mathcal{A}, N_1, N_2$  determine a dynamical system in  $(\alpha_t, \beta_t)$

## Model 1: Mistransmission

- ▶ Individual N, V examples mistransmitted, with some probability.
  - ▶ Mistransmission probabilities for N:

$$a_{21} = P(\text{Hear } \acute{\sigma}\sigma \mid \sigma\acute{\sigma} \text{ intended}), \quad a_{12} = P(\text{Hear } \sigma\acute{\sigma} \mid \acute{\sigma}\sigma \text{ int})$$

$b_{21}, b_{12}$ : same, for V.

- ▶ From examples *heard*, learner **probability matches** for N and V separately:

$$\hat{\alpha} = \frac{k_1}{N_1}, \quad \hat{\beta} = \frac{k_2}{N_2}$$

## Model 1: Properties

Unique fixed point, no bifurcations.

1. Sudden change: ~~X~~
2.  $\{1,1\}$ ,  $\{1,2\}$ ,  $\{2,2\}$ : ~~X~~
3.  $\{2,1\}$ : ~~X~~
4. Observed changes: ~~X~~
5. Frequency dependence: ~~X~~

## Model 2: Prior/data competition (“analogy”)

- ▶ Individual N, V examples **heard correctly**.
- ▶ Estimate probabilities of *grammars*: {1,1}, {1,2}, {2,2}, {2,1}

$$\hat{P}_{11} = \frac{N_1 - k_1}{N_1} \frac{N_2 - k_2}{N_2}, \quad \hat{P}_{12} = \frac{N_1 - k_1}{N_1} \frac{k_2}{N_2}$$
$$\hat{P}_{22} = \frac{k_1}{N_1} \frac{k_2}{N_2}, \quad \hat{P}_{21} = \frac{k_1}{N_1} \frac{N_2 - k_2}{N_2}$$

- ▶ Have prior probabilities for grammars:  $\lambda_{11}, \lambda_{12}, \lambda_{22}, \lambda_{21}$ , **same for all learners**
- ▶ To *produce* N or V:
  - ▶ Choose grammars  $g, g'$  according to  $\vec{P}, \vec{\lambda}$ .
  - ▶ Repeat until  $g = g'$ .
  - ▶ Produce N/V given by  $g$
- ▶ **Assume**  $\lambda_{21} = 0$  ({2,1} not in lexicon).

## Model 2: Properties

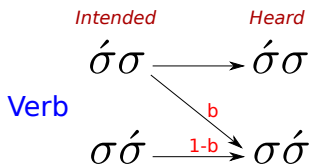
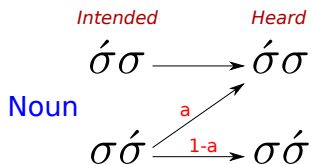
- ▶ 3 fixed points:  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 1)$
- ▶ Bifurcations corresponding to empirically observed changes:
  1.  $\{1, 1\} \rightarrow \{1, 2\}$
  2.  $\{2, 2\} \rightarrow \{1, 2\}$
  3.  $\{1, 2\} \rightarrow \{1, 1\}$
  4.  $\{1, 2\} \rightarrow \{2, 2\}$
- ▶ Change to  $\{1, 2\}$  not frequency-dependent, instead triggered by increasing  $\lambda_{12}$ .

## Model 2: Properties

1. Sudden change: ✓
2.  $\{1,1\}$ ,  $\{1,2\}$ ,  $\{2,2\}$ : ✓
3.  $\ast\{2,1\}$ : ✓
4. Observed changes: ✓
5. Frequency dependence: ✗

### Model 3: Prior/data competition + mistransmission

- ▶ Model 1 + Model 2
- ▶ Get N/V examples, with asymmetric mistransmission:



- ▶ Estimate probs  $\vec{P}$  of grammars:  $\{1,1\}$ ,  $\{1,2\}$ ,  $\{2,2\}$ ,  $\{2,1\}$
- ▶ Have prior probs  $\vec{\lambda}$  for each grammar.
- ▶  $\vec{P}$ ,  $\vec{\lambda}$  compete in production.

## Model 3: Properties

- ▶ Dynamics the same as Model 2, except all changes **frequency-dependent**.
- ▶ In particular, **change to {1,2} triggered by falling frequency**.
- ▶ Keeping  $\vec{\lambda}$  fixed, {2,2} and {1,1} lose stability as frequency decreases.

## Model 3: Properties

- ▶ Dynamics the same as Model 2, except all changes **frequency-dependent**.
- ▶ In particular, **change to {1,2} triggered by falling frequency**.
- ▶ Keeping  $\vec{\lambda}$  fixed, {2,2} and {1,1} lose stability as frequency decreases.

1. Sudden change: ✓
2. {1,1}, {1,2}, {2,2}: ✓
3. \*{2,1}: ✓
4. Observed changes: ✓
5. Frequency dependence: ✓

## Model comparison

	Model		
	1	2	3
Sudden change	✗	✓	✓
{1,1}, {1,2}, {2,2}	✗	✓	✓
*{2,1}	✓	✓	✓
Observed changes	✗	✓	✓
Freq dep	✗	✗	✓

Full model set:

	Model							
	1	2	3	4	5	6	7	8
Sudden change	✗	✓	✓	✗	✗	✗	✗	✓
{1,1}, {1,2}, {2,2}	✗	✓	✓	✗	✗	✓	✗	✓
*{2,1}	✓	✓	✓	✗	✓	✓	✓	✓
Observed changes	✗	✓	✓	✗	✗	✗	✗	✗
Freq dep	✗	✗	✓	✗	✗	✗	✓	✗

Introduction

Data

Causes

Models

**Discussion**

## Model comparison

- ▶ Model 3 works
  - ▶ We are more interested in the many that *don't* work.
  - ▶ Value of exploring the “landscape” of models:
    - ▶ All models plausible a priori.
    - ▶ Most models have **some** desired properties.
    - ▶ Very few have **all** desired properties.
- ⇒ **large model set important** to understand source of observed dynamics, uniqueness of a particular model.
- ▶ Landscape especially valuable for (relatively) detailed data.

## Model comparison

- ▶ Model 3 works
  - ▶ We are more interested in the many that *don't* work.
  - ▶ Value of exploring the “landscape” of models:
    - ▶ All models plausible a priori.
    - ▶ Most models have **some** desired properties.
    - ▶ Very few have **all** desired properties.
- ⇒ **large model set important** to understand source of observed dynamics, uniqueness of a particular model.
- ▶ Landscape especially valuable for (relatively) detailed data.
  - ▶ Previous work on modeling change usually considers 1–2 models.

## Learner biases

- ▶ At broad level, two kinds of bias in models:
  - ▶ Bias in learning data (e.g. mistransmission)
  - ▶ Bias in learner's algorithm (e.g. prior/data competition)
- ▶ In full model set, **only models with *both* kinds of bias give desired dynamics.**
- ▶  $\approx$  correspond to broader, contrasting views on the sources of language change: [Moreton 2008]
  1. "Channel bias": Misperception, misarticulation, processing, frequency..
  2. "Analytic bias": Analogy, markedness, regularization...

## Individual learning, population dynamics

- ▶ In our models, much of interesting behavior comes from the *interaction* of different causes.
  - ▶ Frequency-dependent dynamics not in Model 1 or Model 2, but in Model 3 (=1+2).
- ▶ **Population-level dynamics themselves are a source of change.**

## Individual learning, population dynamics

- ▶ In our models, much of interesting behavior comes from the *interaction* of different causes.
    - ▶ Frequency-dependent dynamics not in Model 1 or Model 2, but in Model 3 (=1+2).
  - ▶ **Population-level dynamics themselves are a source of change.**
  - ▶ Population-level dynamics are all we observe, but may not transparently reflect biases in data, learner.
  - ▶ Given an observed pattern of variation and change, different plausible learning models for individuals give very different population-level dynamics.
- ⇒ Population-level models are important to test any theory of how, why change occurs.

## Thanks

- ▶ Max Bane
- ▶ John Goldsmith
- ▶ Jason Riggle
- ▶ Alan Yu

Dataset, trajectories, slides: [people.cs.uchicago.edu/~morgan](http://people.cs.uchicago.edu/~morgan)

# References 1

- Baker, A. (2008) Computational approaches to the study of language change. *Language and Linguistics Compass*, 2:289–307.
- 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.
- Choudhury, M. (2007) *Computational Models of Real World Phonological Change*. PhD thesis, Indian Institute of Technology Kharagpur.
- 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. & 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.
- Griffiths, T. & Kalish, M. (2007) Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3): 441–480.
- Griffiths, T. & Kirby, S. (2007) Innateness and culture in the evolution of language. *PNAS*, 104(12): 5241–5245.
- Guion, S., Clark, J., Harada, T., and Wayland, R. (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.
- Hudson Kam, C.L. & 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.
- Hudson Kam, C. & Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59:30–66.
- 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.
- Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, M. Studdert-Kennedy, and J. Hurford, J (eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form*. Cambridge: Cambridge University Press. 303–323.

## References 2

- Landsbergen, F. (2009) *Cultural evolutionary modeling of patterns in language change: Exercises in evolutionary linguistics*. PhD thesis, Universiteit Leiden.
- Lightfoot, D. (2001) *How to set parameters: Arguments from language change*. Cambridge: MIT Press.
- Mitchener, W. (2005) Simulating language change in the presence of non-idealized syntax. In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition*. ACL. 10–19.
- Moreton, E. (2008) Analytic bias and phonological typology. *Phonology*, 25:83–127.
- Morgan, J. *From simple input to complex grammar*. Cambridge: MIT Press.
- Mowrey, R. & Pagliuca, W. The reductive character of articulatory evolution. *Rivista di Linguistica*, 7:37–124.
- 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. & 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 & P. Hopper (eds.), *Frequency and the emergence of lexical structure*. John Benjamins, Amsterdam. 137–157.
- Ross, J.R. (1973). Leftward, ho!. In S.R. Anderson & P. Kiparsky (eds.), *A Festschrift for Morris Halle*. 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. & Newport, E.L. (2004) When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, 49:370–407.
- Sonderegger, M. (2010/to appear) Testing for frequency and structural effects in an English stress shift. *BLS* 36.
- Sonderegger, M. & Niyogi, P. (2010/to appear) Dynamical systems models of variation and change: An application to an English stress shift. Ms, University of Chicago.