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:

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,1}</td>
<td>σσ</td>
<td>σσ</td>
<td>anchor, fracture</td>
</tr>
<tr>
<td>{1,2}</td>
<td>σσ</td>
<td>σό</td>
<td>consort, contest</td>
</tr>
<tr>
<td>{2,2}</td>
<td>σό</td>
<td>σό</td>
<td>police, review</td>
</tr>
</tbody>
</table>

- Never {2,1}

- Sherman (1975):
  - Considered N/V pairs, \(\approx 1600-1800\).
  - 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.
  \((1/2=\text{both reported, 1 listed first.})\)

- \(149 \times 62 \times 2 = 18.5k \text{ measurements.}\)

- Allows detailed description of change.
Stress trajectories

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

- **Combat**: The stress levels rise sharply around 1800 and remain high until the late 19th century, showing a significant impact.
- **Converse**: The stress levels show a more gradual increase with sporadic dips, indicating a less immediate but sustained impact over time.
Observed changes

**Common**
{2,2}→{1,2}
{1,1}→{1,2}

**Rarer**
{1,2}→{1,1}
{1,2}→{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\}→\{1,2\} change have lower frequency. [Phillips 1984, Sonderegger 2010/to appear]

- Some evidence that falling frequency (N+V) triggers change:

\begin{center}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Year & 1650 & 1700 & 1750 & 1800 & 1850 & 1900 \\
\hline
N+V Frequency & 1 & 1.2 & 1.4 & 1.6 & 1.8 & 2 \\
\hline
Mvg avg of pron. & 1 & 1.2 & 1.4 & 1.6 & 1.8 & 2 \\
\hline
\end{tabular}
\end{center}

\begin{center}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Year & 1780 & 1800 & 1820 & 1840 & 1860 & 1880 & 1900 \\
\hline
N+V Frequency & 0 & 0.01 & 0.015 & 0.02 & 0.025 & 0.03 & 0.035 \\
\hline
Mvg avg of pron. & 0.01 & 0.015 & 0.02 & 0.025 & 0.03 & 0.035 & 0.04 \\
\hline
\end{tabular}
\end{center}

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).

  - Learners have categoricality bias.

Considered today for N/V case
Sources of change: Mistransmission

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

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 $\rightarrow \dot{o}\sigma$, $V$ biased $\rightarrow \sigma\dot{o}$.

   - $N < V$ (% 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
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]

- System state $\alpha_t$ evolves by $\alpha_{t+1} = f(\alpha_t)$:

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

- Dynamical systems viewpoint: Examine limiting ($t \to \infty$) 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:

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. $^*\{2,1\} \leftrightarrow$ 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 the same number of examples.
  3. Each example is equally likely to come from any member of generation $n - 1$.
  4. Each generation has infinitely many members.

The last assumption is crucial for the dynamics and is necessitated by the existence of variation in speech communities. This differs from "iterated learning" experiments/simulations, where each generation has 1 member. [e.g. Kirby 2000, Kirby et al. 2007, Griffiths & Kalish 2007]
**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]
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:

<table>
<thead>
<tr>
<th>Total</th>
<th>Heard as $\sigma \dot{\sigma}$</th>
<th>Heard as $\dot{\sigma} \sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$ examples</td>
<td>$N_1$</td>
<td>$k_1$</td>
</tr>
<tr>
<td>$V$ examples</td>
<td>$N_2$</td>
<td>$k_2$</td>
</tr>
</tbody>
</table>

- Applies $\mathcal{A}$ to learn:

  $\hat{\alpha}$: Prob of producing $N$ as $\sigma \dot{\sigma}$

  $\hat{\beta}$: ””

- $\alpha_t$, $\beta_t$:
  Probability random $N$, $V$ example from generation $t$ produced as $\sigma \dot{\sigma}$. (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:
    \[ a_{21} = P(\text{Hear } \sigma \sigma | \sigma \sigma \text{ intended}), \quad a_{12} = P(\text{Hear } \sigma \sigma | \sigma \sigma \text{ int}) \]
    \[ b_{21}, b_{12}: \text{ same, for V.} \]
  - 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: ✗
2. \{1,1\}, \{1,2\}, \{2,2\}: ✗
3. *\{2,1\}: ✗
4. Observed changes: ✗
5. Frequency dependence: ✗
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. \{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:

```
\[ \hat{\sigma} \sigma \rightarrow \hat{\sigma} \sigma \]
\[ \sigma \hat{\sigma} \rightarrow \sigma \hat{\sigma} \]
```

- Estimate probs $\tilde{P}$ of grammars: $\{1,1\}$, $\{1,2\}$, $\{2,2\}$, $\{2,1\}$
- Have prior probs $\tilde{\lambda}$ for each grammar.
- $\tilde{P}$, $\tilde{\lambda}$ 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

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Sudden change</td>
<td>✗</td>
</tr>
<tr>
<td>{1,1}, {1,2}, {2,2}</td>
<td>✗</td>
</tr>
<tr>
<td>*{2,1}</td>
<td>✓</td>
</tr>
<tr>
<td>Observed changes</td>
<td>✗</td>
</tr>
<tr>
<td>Freq dep</td>
<td>✗</td>
</tr>
</tbody>
</table>
Full model set:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sudden change</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>{1,1}, {1,2}, {2,2}</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>*{2,1}</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Observed changes</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Freq dep</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
</tbody>
</table>
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.

$\Rightarrow$ 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.

- ≈ 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...
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
References 1


