EMERGENCE OF LEXICAL IDIOSYNCRASY IN LANGUAGE CHANGE:
An iterated learning simulation

Claire Moore-Cantwell
Simon Fraser University
Introduction

Across languages, more frequent lexical items diverge more from the grammar:

English Past Tense: irregulars more frequent than regulars

Bybee, 1995: Higher frequency words have greater “autonomy”

Morgan and Levy, 2016: Experience → Idiosyncrasy and autonomy from the grammar
Introduction

Across languages, more frequent lexical items diverge more from the grammar.

*Today:* Modeling divergence from gradient phonology

- **Representational Strength Theory**
  
  *Gradient memory strength for properties of lexical items*

- The **Gradient Lexicon and Phonology Learner (GLaPL)**
  
  - Integrates learning of lexicon and probabilistic phonology
  - (Phonology affects lexical storage: predictable properties not stored)
  - Frequency affects lexical storage: exposure → more detailed representations
  - Over time, detailed representations → exceptions
Frequency and exceptionality

*Higher frequency → More idiosyncratic*

**English Comparative:** words vary between *more* and *–er*

- happier ~ more happy
- bigger ~ ?? more big

More frequent → more categorical
Less frequent → grammar determines output

- monoyllables → *–er*
- final r/l → *more*
  ...

*Boyd, 2012; Smith and Moore-Cantwell, 2017*
Frequency and exceptionality

Higher frequency → More idiosyncratic

**English Binomial Expressions:** conjuncts vary in order

- lemons and cucumbers ~ cucumbers and lemons
- bread and butter ~ ?? butter and bread

shorter first
more powerful first
  (bishops and priests)

...
Frequency and exceptionality

Higher frequency → More idiosyncratic

**English Binomial Expressions:** conjuncts vary in order

- lemons and cucumbers ~ cucumbers and lemons
- bread and butter ~ ?? butter and bread

| shorter first                                |
|                                             |
| more powerful first                         |
| (bishops and priests)                       |
| ...                                          |

*Morgan & Levy, 2015, 2016*
Frequency and exceptionality

*Higher frequency ➞ More idiosyncratic*

**Subject Pronouns in Spanish:** Subject pronouns are optional

*Hablo ~ Yo hablo*
*Digo ~ ?? Yo digo*

*Erker & Guy, 2012*
Frequency and exceptionality

Higher frequency → More idiosyncratic

In patterns of within-item variation:

Higher frequency forms:

- Diverge from the predictions of the variable grammar
- Exhibit more extreme behavior, varying less as an item than their low-frequency counterparts

Experience → autonomy from the grammar, consistency
Frequency and exceptionality

Higher frequency $\rightarrow$ More idiosyncratic

MaxEnt grammar model
+ learning/representation of words’ features
  
  Representational Strength Theory
+ learning algorithm for both
  
  Gradient Lexicon and Phonology Learner (GLaPL)

Iterated learning (output of learning is input to next “generation”)
  $\rightarrow$ High-frequency items in variable patterns become extreme
Modeling probabilistic generalizations

Constraints conflict, and determine a probability distribution over output candidates

<table>
<thead>
<tr>
<th></th>
<th>( p )</th>
<th>( H )</th>
<th>OCP-LIQ</th>
<th>( \sigma )-ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>foul + Comp</td>
<td></td>
<td></td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>( \rightarrow ) more foul</td>
<td>0.59</td>
<td>-1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>( \rightarrow ) fouler</td>
<td>0.41</td>
<td>-1.4</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ H = - \sum W_i \ast V_i \]

“Harmony”

(Smolensky and Legendre, 2006; Pater, 2016)

\[ p = \frac{e^H}{\sum e^H} \]

\( H \)armony

(Maximum Entropy Grammar)

(Goldwater and Johnson, 2003)

Predicts intra-speaker variation

For a given speaker, \( p \) is the probability that they will produce that output on any given utterance of the input word.
Adding in word knowledge

What to do with higher-frequency words that don’t follow the grammar?

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>( H )</th>
<th>OCP-LIQ</th>
<th>( \sigma )-ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>small + ( \text{Comp} )</td>
<td></td>
<td></td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>( \times \rightarrow ) more small</td>
<td>0.59</td>
<td>-1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>99.6% ( \checkmark \rightarrow ) smaller</td>
<td>0.41</td>
<td>-1.4</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Speakers must memorize the behavior of words like \textit{small} + \textit{Comp}
Adding in word knowledge

Proposal: Representational Strength Theory  (compare: Direct OT Golston, 1996)

Phonological Form Constraints (PFC’s)

-er – SMALL: Assign a violation to any output form for the input SMALL which also contains a + COMP, and does not use the suffix –er to express it

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>H</th>
<th>-er</th>
<th>OCP-LIQ</th>
<th>σ-ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALL + COMP</td>
<td></td>
<td></td>
<td>5.4</td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>more small</td>
<td>0</td>
<td>-6.4</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>smaller</td>
<td>0.99</td>
<td>-1.4</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SMALL 5.4 1st segment sibilant
7.2 2nd segment labial
6.7 3rd segment low
7.9 3rd segment voiced
...
Adding in word knowledge

Proposal: Representational Strength Theory with Phonological Form Constraints

<table>
<thead>
<tr>
<th>Word</th>
<th>-er</th>
<th>Pos1 SIBILANT</th>
<th>Pos1 CORONAL</th>
<th>Pos2 NASAL</th>
<th>Pos3 ALVEOLAR</th>
<th>Pos3 VOICE</th>
<th>Pos1 RHOTIC</th>
<th>Pos1 NASAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small + Comp</td>
<td>5.4</td>
<td>6.2</td>
<td>5.8</td>
<td>7.4</td>
<td>2.5</td>
<td>0.7</td>
<td>7.2</td>
<td>5.6</td>
</tr>
<tr>
<td>rm small</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tmalə</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fmalə</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spalə</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smələ</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smələ</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smale</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>prəd small</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>→ smaə</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

No Underlying Form!
No Faithfulness constraints

PFC's are the phonological part of the lexical entry (compare: Direct OT Golston, 1996)

Gradient weight ~ gradient memory resource allocation

13
Markedness can overcome PFCs

Proposal: Representational Strength Theory w/ Phonological Form Constraints

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>$H$</th>
<th>$^{*}\mathrm{t}\check{\mathrm{v}}$</th>
<th>Pos4 +stop</th>
<th>Pos4 +cor</th>
<th>...</th>
<th>Pos1 +high</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREET + PROG</td>
<td></td>
<td></td>
<td>10</td>
<td>5</td>
<td>10</td>
<td></td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>$\rightarrow$ grírinya</td>
<td>0.99</td>
<td>-5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grítinya</td>
<td>0</td>
<td>-10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grípinya</td>
<td>0</td>
<td>-10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gríreña</td>
<td>0</td>
<td>-13</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Next: Learning weights of Markedness and PFC’s...
Learning probabilistic generalizations

*Error Driven Learning* (Boersma and Hayes, 2001; Rosenblatt, 1958)

- Guess values for constraint weights
- Use guess to predict output for a word
- Check prediction against observed output
- Adjust ‘guess’
- Do nothing

Learns one word at a time

**Starting guess:**
All weights zero

**Each learning step:**
Sample a word based on frequency
Learning probabilistic generalizations

*Error Driven Learning* (Boersma and Hayes, 2001; Rosenblatt, 1958)

Sample $t$: *smaller*

Use current state of grammar to predict correct output:

<table>
<thead>
<tr>
<th></th>
<th>$p$</th>
<th>$\mathcal{H}$</th>
<th>OCP-LIQ</th>
<th>$\sigma$-ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALL + COMP</td>
<td></td>
<td></td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>more small</td>
<td>0.59</td>
<td>-1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>smaller</td>
<td>0.41</td>
<td>-1.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Randomly sample: *more small*

Does not match observed pronunciation!

Update weights

$\Delta w = 0.01$

OCP-LIQ favors the incorrect outcome *decrease*

$\sigma$-ER favors the correct outcome *increase*

Weights only change a little at a time
Learning probabilistic generalizations

*Error Driven Learning* (Boersma and Hayes, 2001; Rosenblatt, 1958)

Sample $t$: smaller

Use current state of grammar to predict correct output:

<table>
<thead>
<tr>
<th></th>
<th>$p$</th>
<th>$H$</th>
<th>OCP-LIQ</th>
<th>$\sigma$-ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALL + COMP</td>
<td></td>
<td></td>
<td>1.39</td>
<td>1.01</td>
</tr>
<tr>
<td>more small</td>
<td>0.58</td>
<td>-1.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>smaller</td>
<td>0.42</td>
<td>-1.39</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

randomly sample: more small

Does not match observed pronunciation!

Weights only change a little at a time $\Delta w = 0.01$

$OCP$-$LIQ$ favors the incorrect outcome *decrease*

$\sigma$-$ER$ favors the correct outcome *increase*
Adding in word learning
The Gradient Lexicon and Phonology Learner (GLaPL)

Error Driven Learning (Boersma and Hayes, 2001; Rosenblatt, 1958)

- Guess values for constraint weights
- Decay PFC weights Use guess to predict output for a word
- Learn one word at a time
- Check prediction against observed output
- Adjust weights
  - Induce a Phonological Form Constraint
  - starting weight: 10
- Do nothing
  - mismatch (error)
  - match

Decay PFC weights
Use guess to predict output for a word

Learns one word at a time
Learning probabilistic generalizations

*Error Driven Learning* (Boersma and Hayes, 2001; Rosenblatt, 1958)

Sample $t$: *smaller*

Use current state of grammar to predict correct output:

<table>
<thead>
<tr>
<th>Sample</th>
<th>$p$</th>
<th>$H$</th>
<th>-er</th>
<th>OCP-LIQ</th>
<th>σ-ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALL + COMP</td>
<td></td>
<td></td>
<td>-er</td>
<td></td>
<td></td>
</tr>
<tr>
<td>more small</td>
<td>0</td>
<td>-11.01</td>
<td>1</td>
<td>1.39</td>
<td>1.01</td>
</tr>
<tr>
<td>smaller</td>
<td>0.99</td>
<td>-1.39</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

randomly sample:

Does not match observed pronunciation!

Update weights

*Induce Phonological Form Constraint*
Decay

- Phonological Form Constraints (PFC’s) = memory for correct pronunciation of the word

- Elements of declarative memory decay over time (Hintzmann, 1984; Brady et al., 2013)
  - All PFC’s decay at the same rate ($10^{-4}$)
  - Decay to zero $\rightarrow$ removed from consideration
    
    But could be added back later
Frequency and exceptionality

Gradient Lexicon and Phonology Learner (GLaPL)

1000 words, 5 exceptions:

![Graph showing weight of PFCs over iterations](image-url)
Frequency and exceptionality

Gradient Lexicon and Phonology Learner (GLaPL)

Fewer, lower weighted PFC’s on low-frequency words
Frequency and exceptionality
Gradient Lexicon and Phonology Learner (GLaPL)

Training data:
• Comparatives in COCA: 4600 adjectives, 1.1 million instances
  (Smith and Moore-Cantwell, 2017)

Constraints:
• One for each phonological conditioning factor
  (Word length, final l/r, stress pattern...)

Parameters: (summary)
5 million learning iterations
Markedness constraints updated by learning rate: 0.01
PFC starting weight: 10
PFC learning rate: 0.1
PFC decay rate: 0.0001
COCA
(observed probabilities)
COCA
(observed probabilities)

GLaPL
(predicted probabilities)

Higher frequency → More idiosyncratic
Lower frequency → Reliance on grammar

1 syllable, -CC
2 syllables, -r
3+ syllables
GLaPL: Exceptionality over generations

Starting state: 1000 toy words: All 50% *more*, 50% *-er*

Words’ frequencies in Zipfian distribution (like natural languages)

Each generation learns, then final state becomes input to next generation (iterated learning)

Griffiths and Kalish, 2007
Kirby et al 2014, ...

26
GLaPL: Exceptionality over generations

Starting state  1000 toy words: All 50% more, 50% -er

Two (relatively dumb) markedness constraints: Be more, Be -er

Parameters: (summary)
500,000 learning iterations
Markedness constraints updated by learning rate: 0.01
PFC starting weight: 10
PFC learning rate: 0.1
PFC decay rate: 0.0001
Input data

Generation 1

Generation 2

Generation 3

Generation 5

Generation 20

Generation 50

log(f) < 3

log(f) > 5
Consistency across runs

Generation 20: Highest density point is always close to 1 or 0 for high-frequency words, and always middling for low-frequency words.

All runs get the basic pattern: high-frequency words are idiosyncratic, while low-frequency words vary according to the grammar.
Conclusions

Frequency is tied to divergence from the Phonological Grammar:

This model (GLaPL) uses:

- Maximum Entropy Grammar model of phonology
- Error-driven learning algorithm
- Phonological Form Constraints: induced on error, and decay over time

- Frequency affects lexical storage: exposure $\rightarrow$ more detailed representations
- Over time, detailed representations $\rightarrow$ exceptions
Thank you!

github.com/clairemoorecantwell/GLaPL
Morphological Composition with Representational Strength Theory

**PERSON**
- Pos1: -voice 8
- Pos1: +stop 6.2
- Pos2: +voice 3.1
- Pos3: +sibilant 7.7
- Pos3: -voice 6
- Initial stress 10

**PL**
- Pos1: -voice 8
- Pos1: +sibilant 6.2
- Pos1: +cont 3.1
- Initial stress 10

Choose between the stored version and the composed version however you want.

### Table

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>f</th>
<th>*[w0i][+voi]</th>
<th>Pos1</th>
<th>Pos2</th>
<th>Pos3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>pl</td>
<td>0.99</td>
<td>-0.2</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>prsnz</td>
<td>0</td>
<td>-9.3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>prsns</td>
<td>0</td>
<td>-13.1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Markedness can overcome PFCs

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>H</th>
<th>*VTv 10</th>
<th>Pos4 +stop 5</th>
<th>Pos4 +cor 10</th>
<th>...</th>
<th>Pos1 +high 8</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GREET + PROG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grírĩŋ</td>
<td>0.99</td>
<td>-5</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grítĩŋ</td>
<td>0</td>
<td>-10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grípĩŋ</td>
<td>0</td>
<td>-10</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grírɛŋ</td>
<td>0</td>
<td>-13</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GLaPL trying to learn crazy data
GLaPL trying to learn crazy data
French schwa alternations

French Schwa alternations

semaine ~ smaine
semetre ~ smestre

Data from Racine, 2007

12 Native speakers rated 2189 nouns with and without schwa