EMERGENCE OF LEXICAL IDIOSYNCRASY IN LANGUAGE CHANGE:

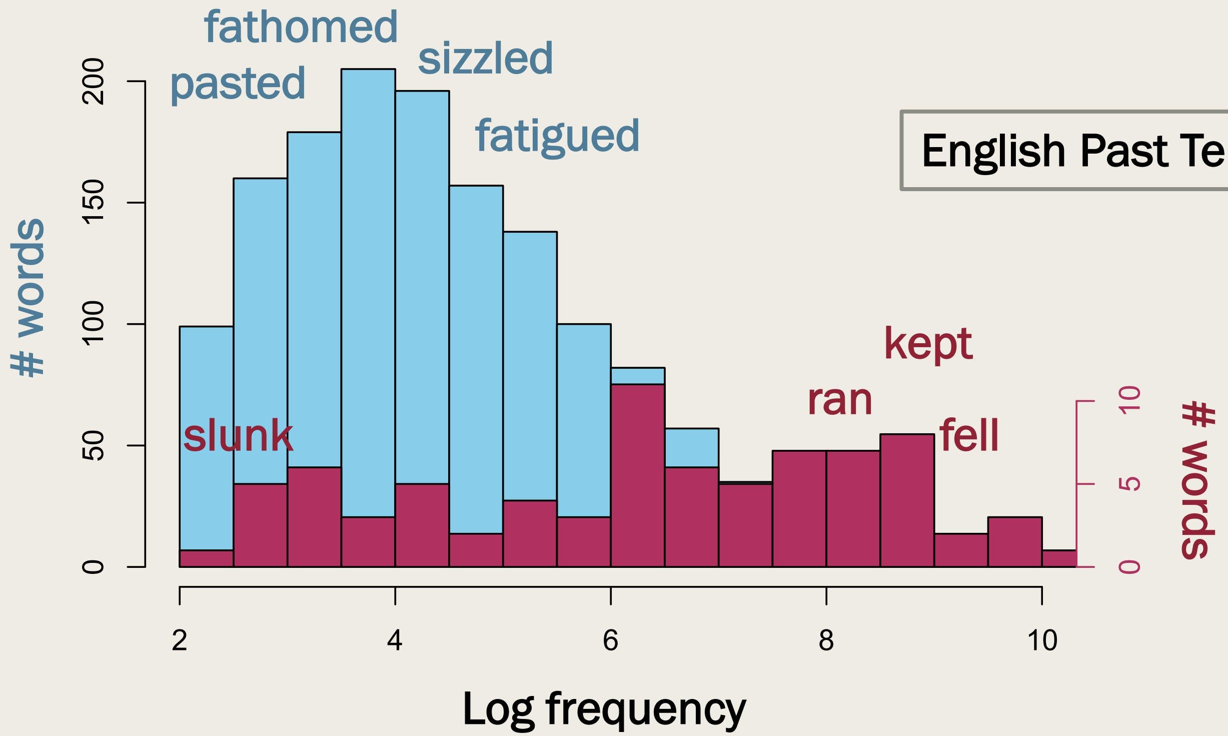
An iterated learning simulation

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 **G**radient **L**exicon **a**nd **P**honology **L**earner (**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

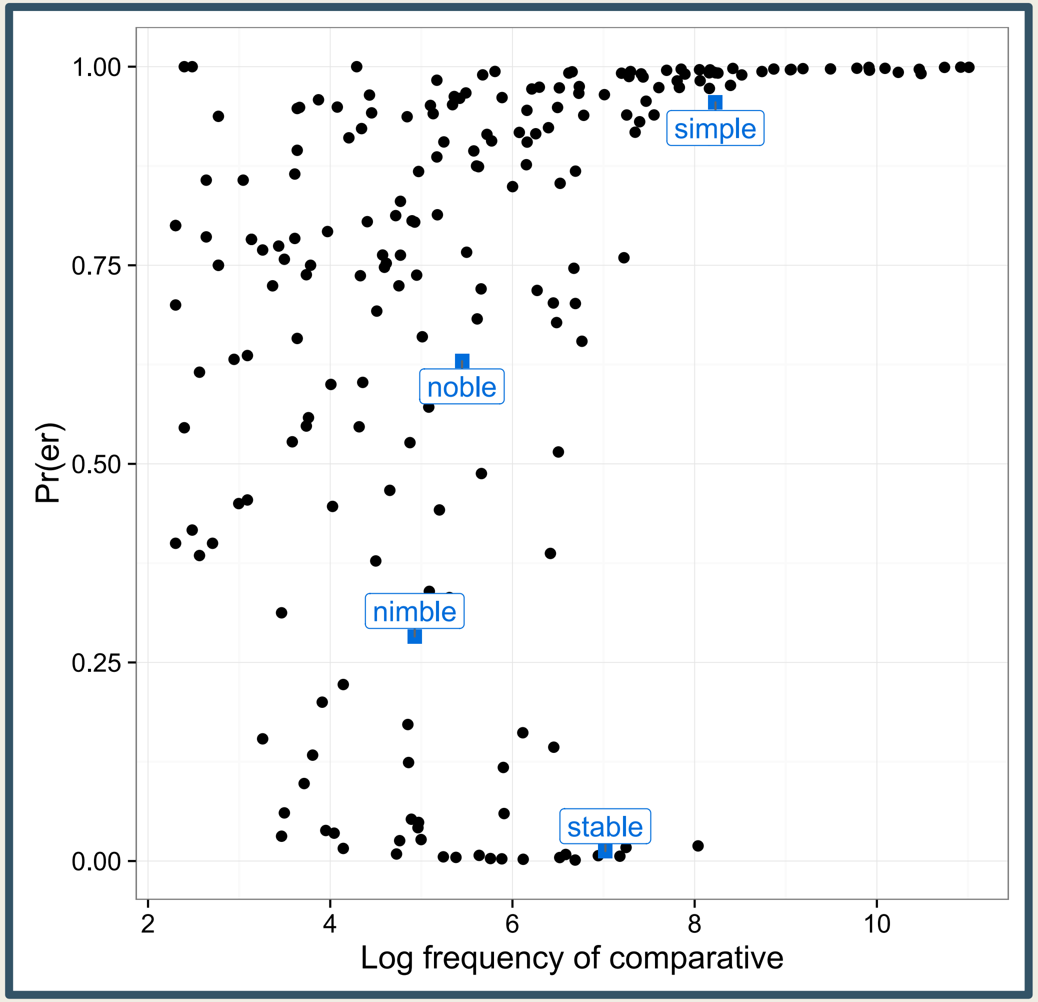
English Comparative: words vary between *more* and *–er*

happier ~ more happy
bigger ~ ?? more big

More frequent → more categorical
Less frequent → grammar determines output

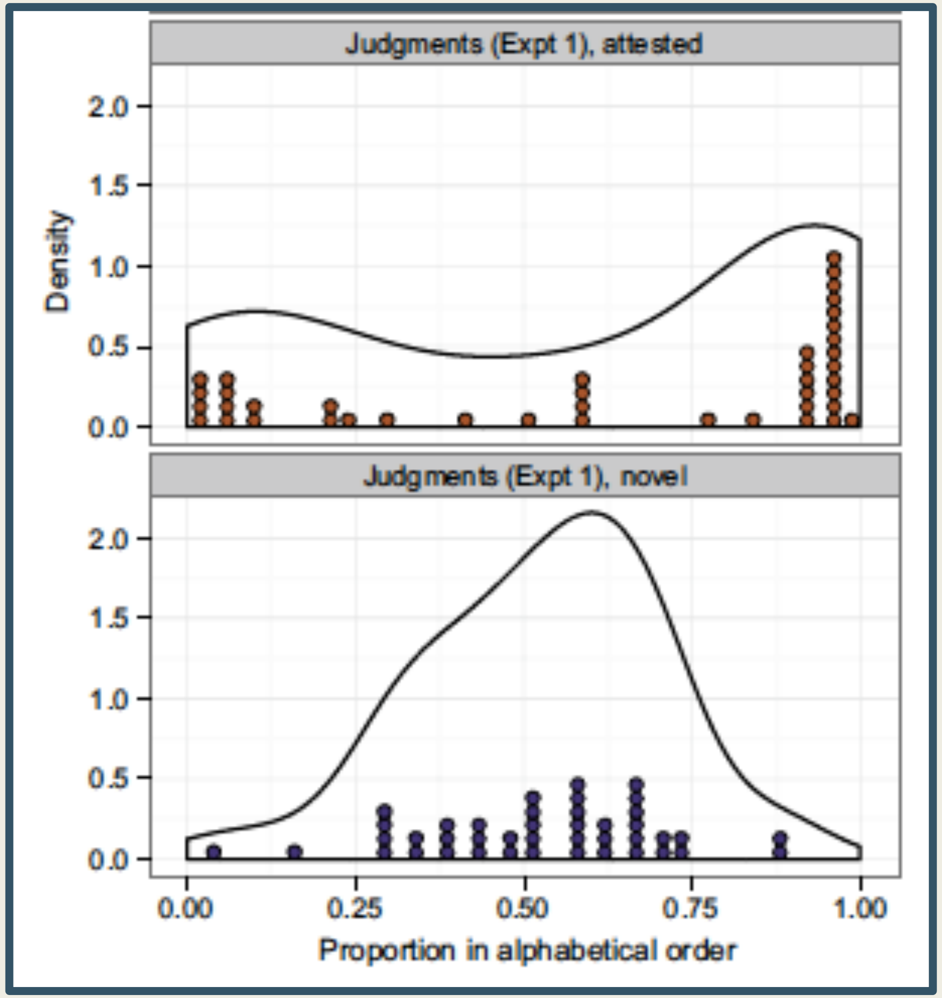
```
monoyllables → -er
final r/l → more
...
```

Boyd, 2012; Smith and Moore-Cantwell, 2017



Frequency and exceptionality

Higher frequency → More idiosyncratic



English Binomial Expressions: conjuncts vary in order

lemons and cucumbers ~ cucmbers and lemons
bread and butter ~ ?? butter and bread

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

Morgan & Levy, 2015, 2016

Frequency and exceptionality

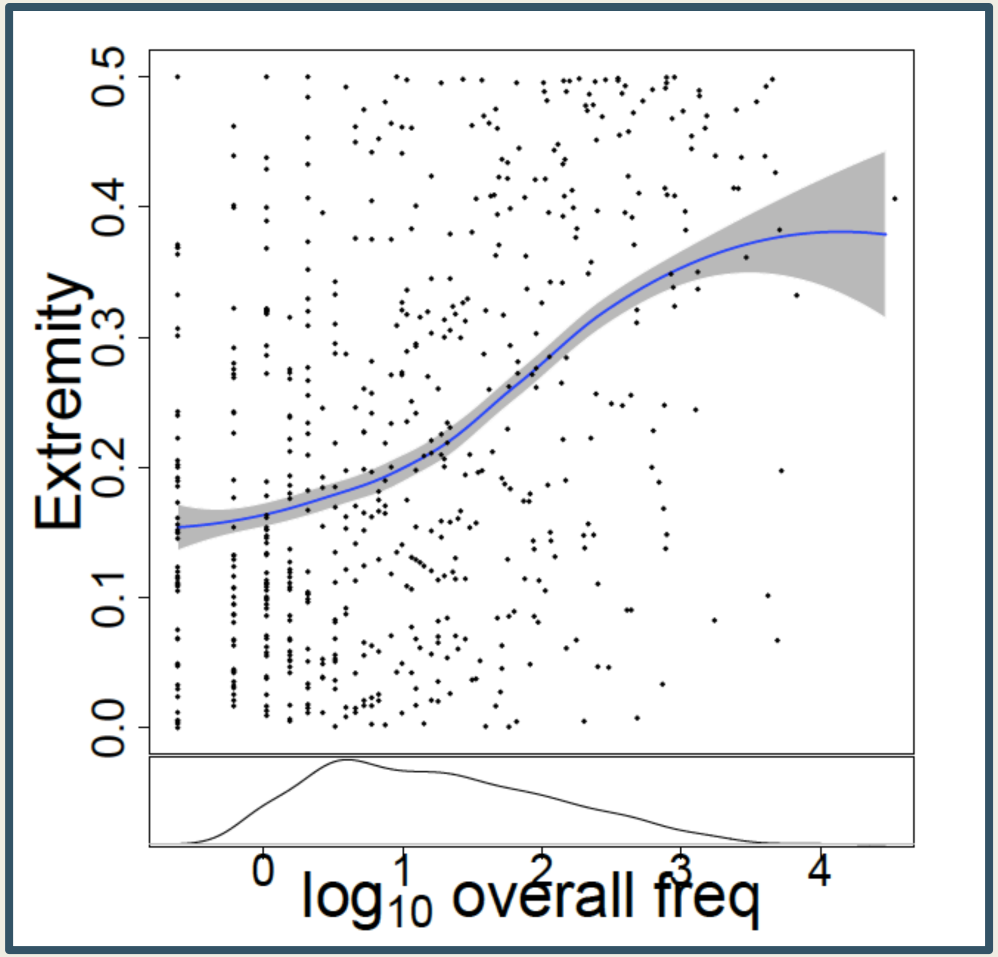
Higher frequency → More idiosyncratic

English Binomial Expressions: conjuncts vary in order

lemons and cucumbers ~ cucmbers 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

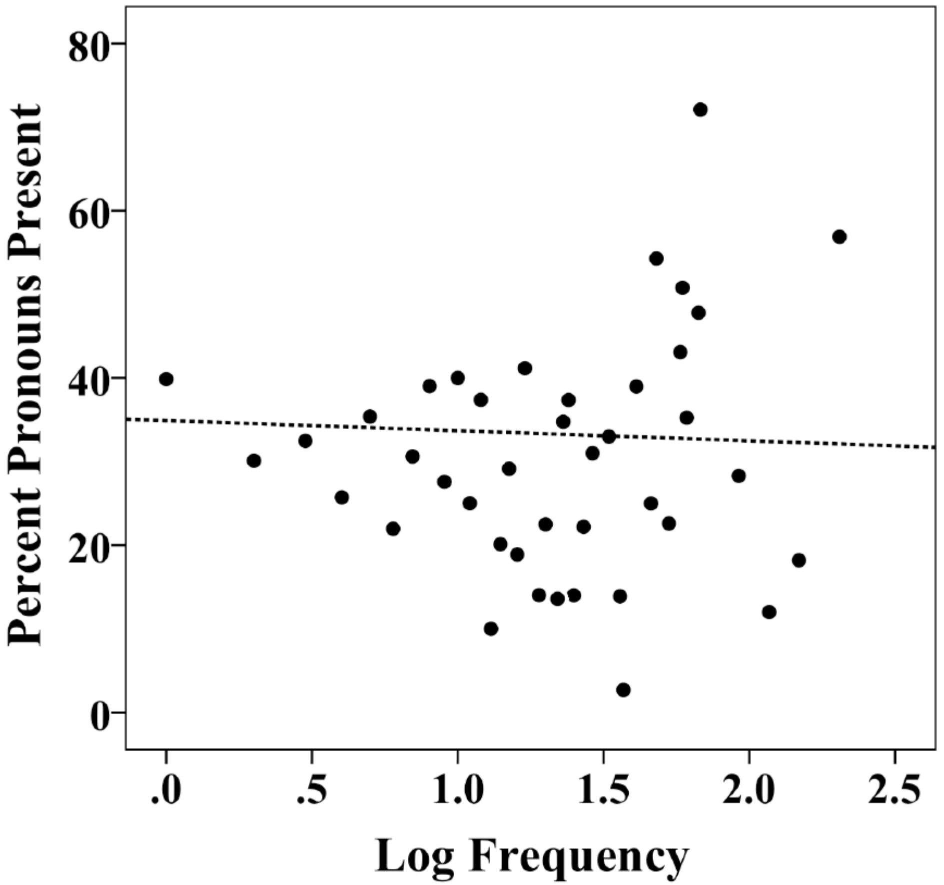
Subject Pronouns in Spanish: Subject pronouns are optional

Hablo ~ Yo hablo
Digo ~ ?? Yo digo

```
Tense-Mood-Aspect
Switch Reference
...
```

Erker & Guy, 2012

Figure 14. Log frequency and percent SPPs present



Frequency and exceptionality

Higher frequency → More idiosyncratic

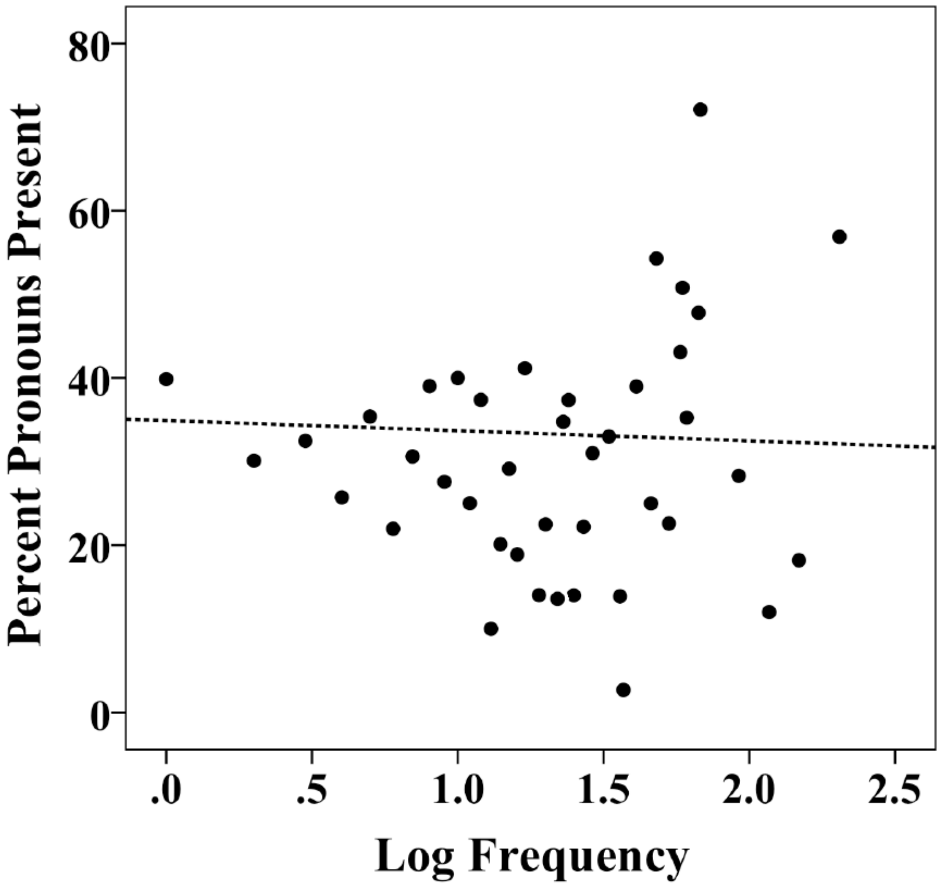
In patterns of within-item variation:

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

Figure 14. Log frequency and percent SPPs present



Frequency and exceptionality

Higher frequency → More idiosyncratic

MaxEnt grammar model

+ *learning/representation of words' features*

Representational Strength Theory

+ *learning algorithm for both*

Gradient **L**exicon **a**nd **P**honology **L**earner (**GLaPL**)

Iterated learning (output of learning is input to next "generation")

→ High-frequency items in variable patterns become extreme

Modeling probabilistic generalizations

Constraints conflict, and determine a probability distribution over output candidates

	p	$\mathcal{H}$	OCP-LIQ 1.4	σ-ER 1
foul + COMP				
→ more foul	0.59	-1		1
→ fouler	0.41	-1.4	1	

MAXIMUM ENTROPY GRAMMAR
(Goldwater and Johnson, 2003)

$$\underbrace{\mathcal{H} = -\sum W_i * Vi}_{\text{“Harmony”}}$$

$$p = \frac{e^{\mathcal{H}}}{\sum e^{\mathcal{H}}}$$

“Harmony”
(Smolensky and Legendre, 2006; Pater, 2016)

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?

		p	$\mathcal{H}$	OCP-LIQ 1.4	σ-ER 1
small + COMP					
	✗ → more small	0.59	-1		1
99.6%	✓ → smaller	0.41	-1.4	1	

Speakers must memorize the behavior of words like *small* + *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

	p	$\mathcal{H}$	*-er* 5.4	OCP-LIQ 1.4	σ-ER 1
SMALL + COMP					
more small	0	-6.4	1		1
→ smaller	0.99	-1.4		1	

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

No Underlying Form!
No Faithfulness constraints

Proposal: Representational Strength Theory w/ **P**honological **F**orm **C**onstraints

	-er 5.4	Pos1 +SIBILANT 6.2	Pos1 +CORONAL 5.8	Pos2 +NASAL 7.4	Pos3 +ALVEOLAR 2.5	Pos3 +VOICE 0.7	Pos1 +RHOTIC 7.2	Pos1 +NASAL 5.6	…
SMALL + COMP									
mɔɹ smal	1								
tmalɚ		1							
ʃmalɚ			1						
spalɚ				1					
smɛlɚ					1				
smal̥ɚ						1			
smalə							1		
pɔɹ smal	1							1	
→ smalɚ									

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

Gradient weight ~ gradient memory resource allocation

Markedness can overcome PFCs

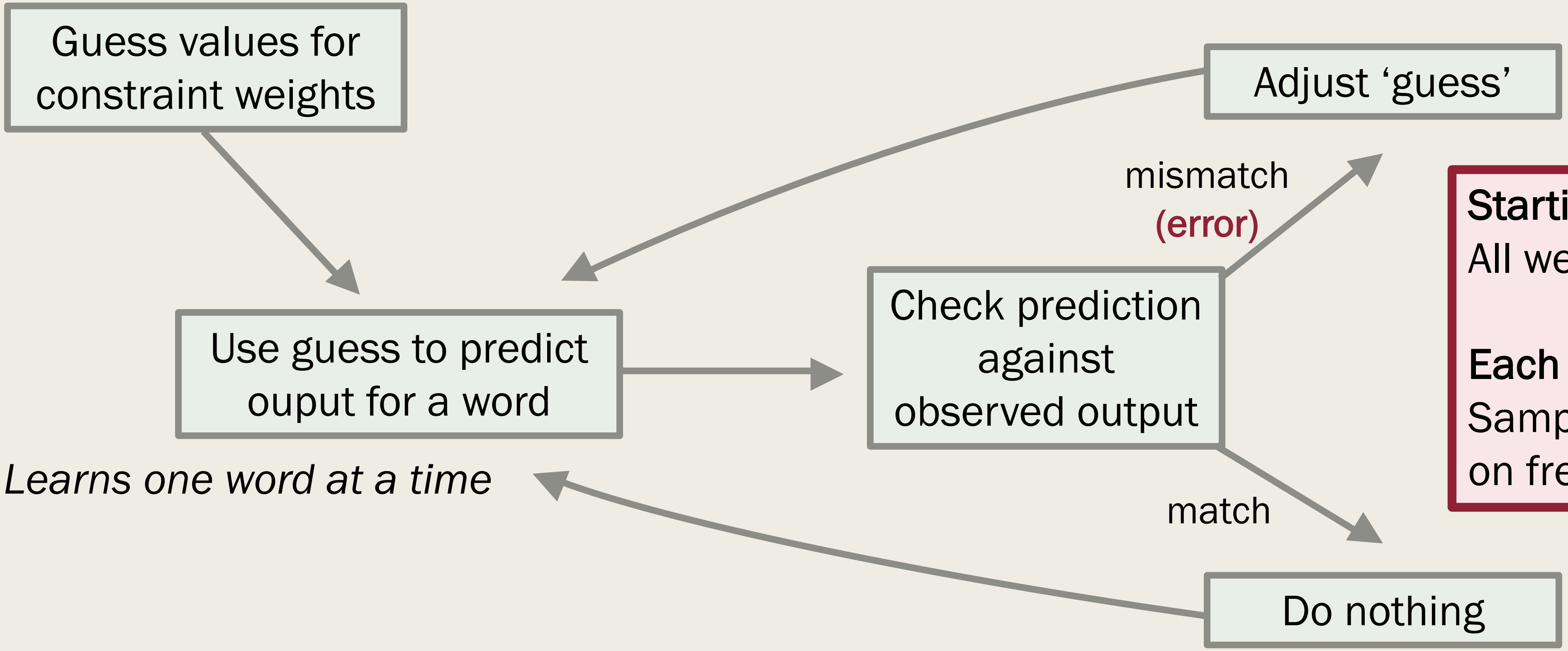
Proposal: Representational Strength Theory w/ Phonological Form Constraints

	p	$\mathcal{H}$	*V́tV̆ 10	*Pos4 +stop 5*	*Pos4 +cor 10*	…	Pos1 +high 8	…
GREET + PROG								
→ gríɾɪŋ	0.99	-5		1				
grítɪŋ	0	-10	1					
grípɪŋ	0	-10			1			
gríɾəŋ	0	-13		1			1	

Next: Learning weights of Markednes and PFC's…

Learning probabilistic generalizations

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



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:

	p	$\mathcal{H}$	OCP-LIQ 1.4 ↓	σ-ER 1 ↑
SMALL + COMP				
more small	0.59	-1		1
smaller	0.41	-1.4	1	

randomly sample: more small

Does not match observed pronunciation! → Update weights

OCP-LIQ favors the incorrect outcome *decrease*

σ-ER favors the correct outcome *increase*

Weights only change a little at a time

$\Delta w = 0.01$

Learning probabilistic generalizations

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

Sample *t:* smaller

Use current state of grammar to predict correct output:

	p	$\mathcal{H}$	OCP-LIQ ↓ 1.39	σ-ER ↑ 1.01
SMALL + COMP				
more small	0.58	-1.01		1
smaller	0.42	-1.39	1	

randomly sample:

more small

Does not match observed pronunciation! → Update weights

OCP-LIQ favors the incorrect outcome
decrease

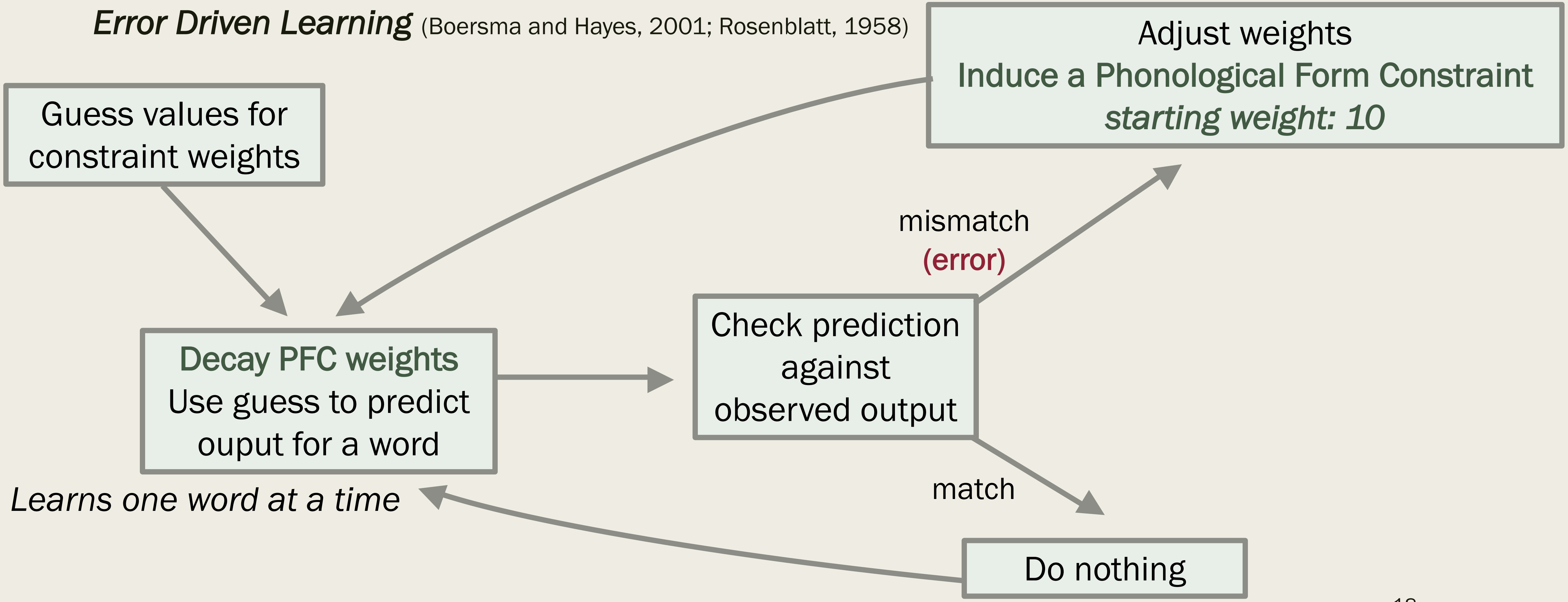
σ-ER favors the correct outcome
increase

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

Adding in word learning

The Gradient Lexicon and Phonology Learner (GLaPL)

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



Learning probabilistic generalizations

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

Sample *t:* smaller

Use current state of grammar to predict correct output:

	p	$\mathcal{H}$	*-er* 10	OCP-LIQ 1.39 ↓	σ-ER 1.01 ↑
SMALL + COMP					
more small	0	-11.01	1		1
smaller	0.99	-1.39		1	

randomly sample:

more small

Does not match observed pronunciation!

Update weights
Induce Phonological Form Constraint

Decay

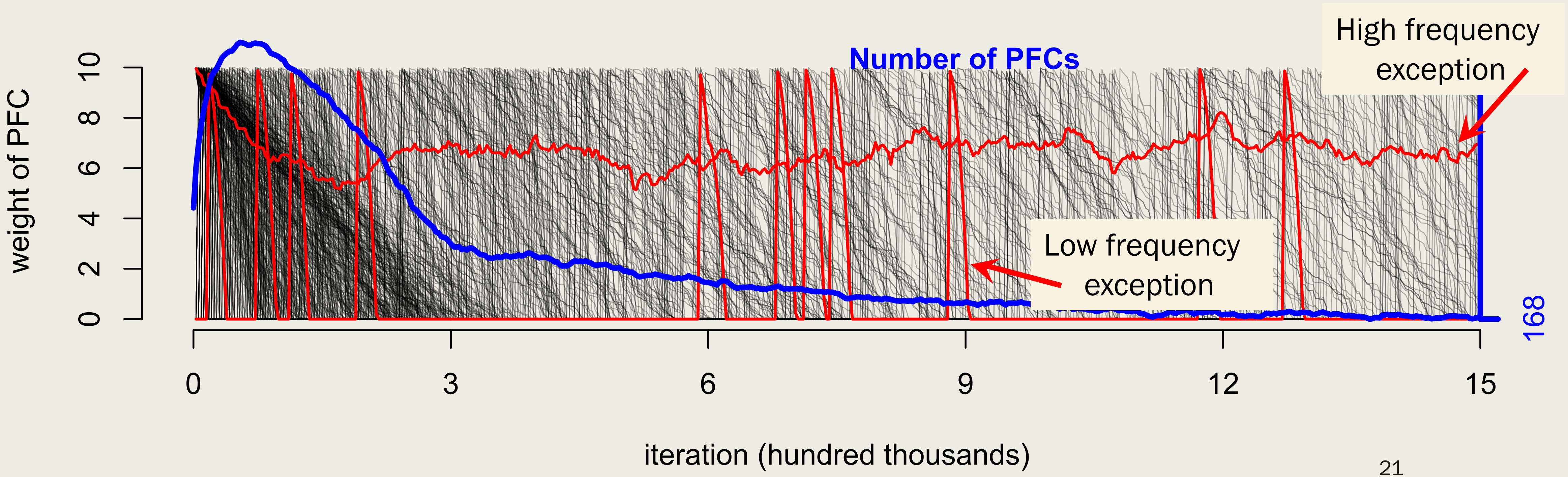
- Phonological Form Constraints (PFC's) = memory for correct pronounciation of the word
- Elements of declarative memory decay over time (Hintzmann, 1984; Brady et al., 2013)
  - *All PFC's decay at the same rate (10^{-4})*
  - *Decay to zero → removed from consideration*

    But could be added back later

Frequency and exceptionality

Gradient Lexicon and Phonology Learner (GLaPL)

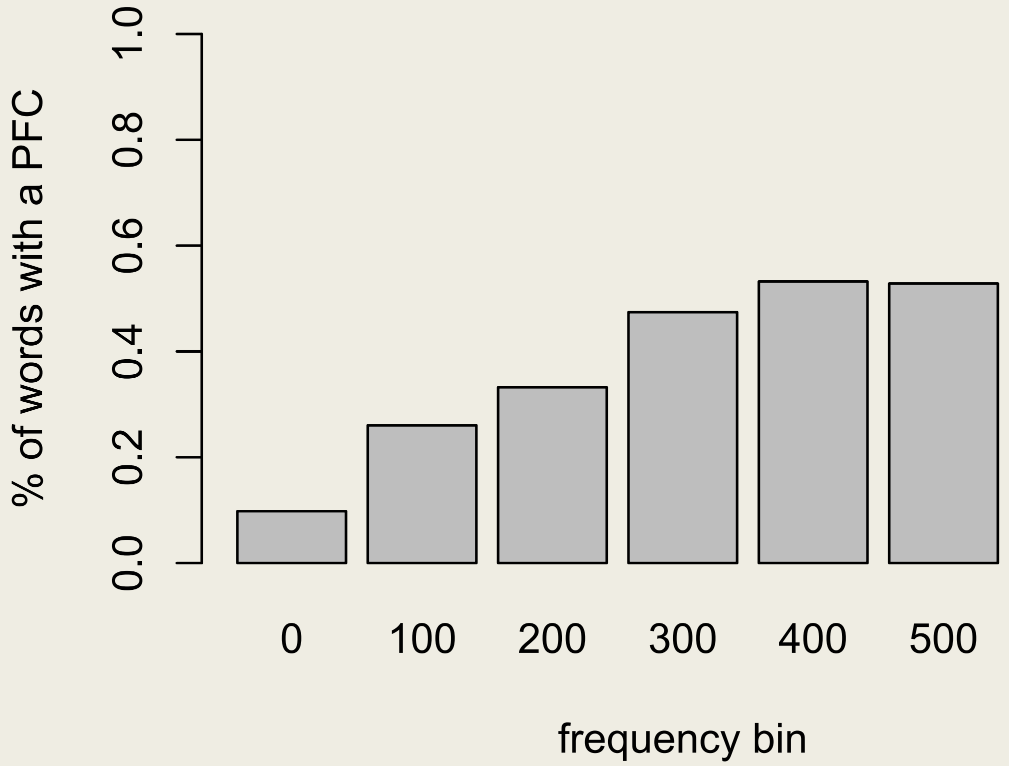
1000 words, 5 exceptions:

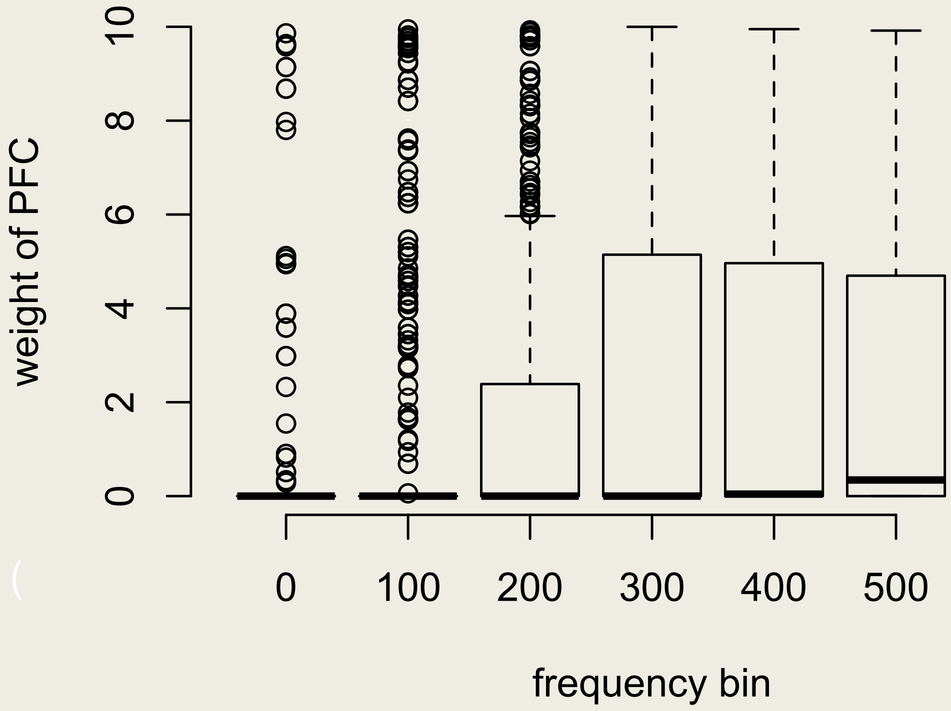


Frequency and exceptionality

Gradient Lexicon and Phonology Learner (GLaPL)

Fewer, lower weighted PFC's on low-frequency words





Frequency and exceptionality

Gradient **L**exicon **a**nd **P**honology **L**earner (**GLaPL**)

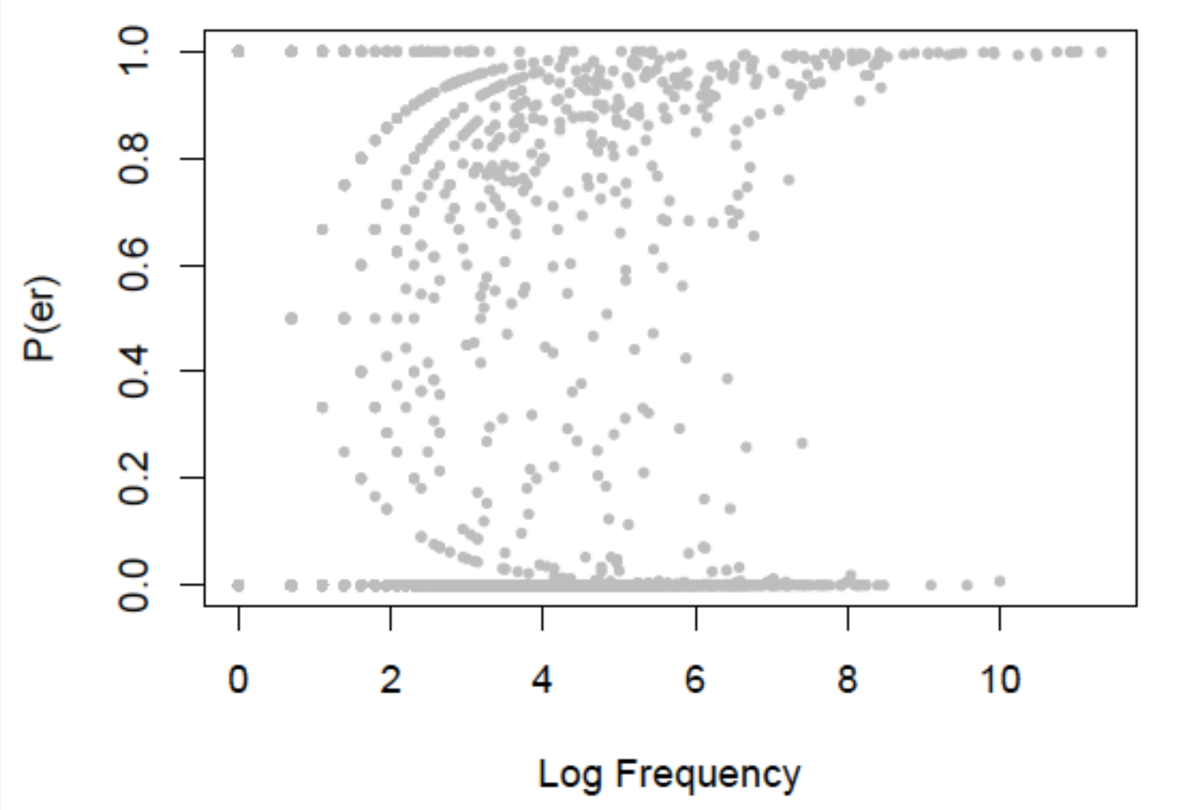
Training data:

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

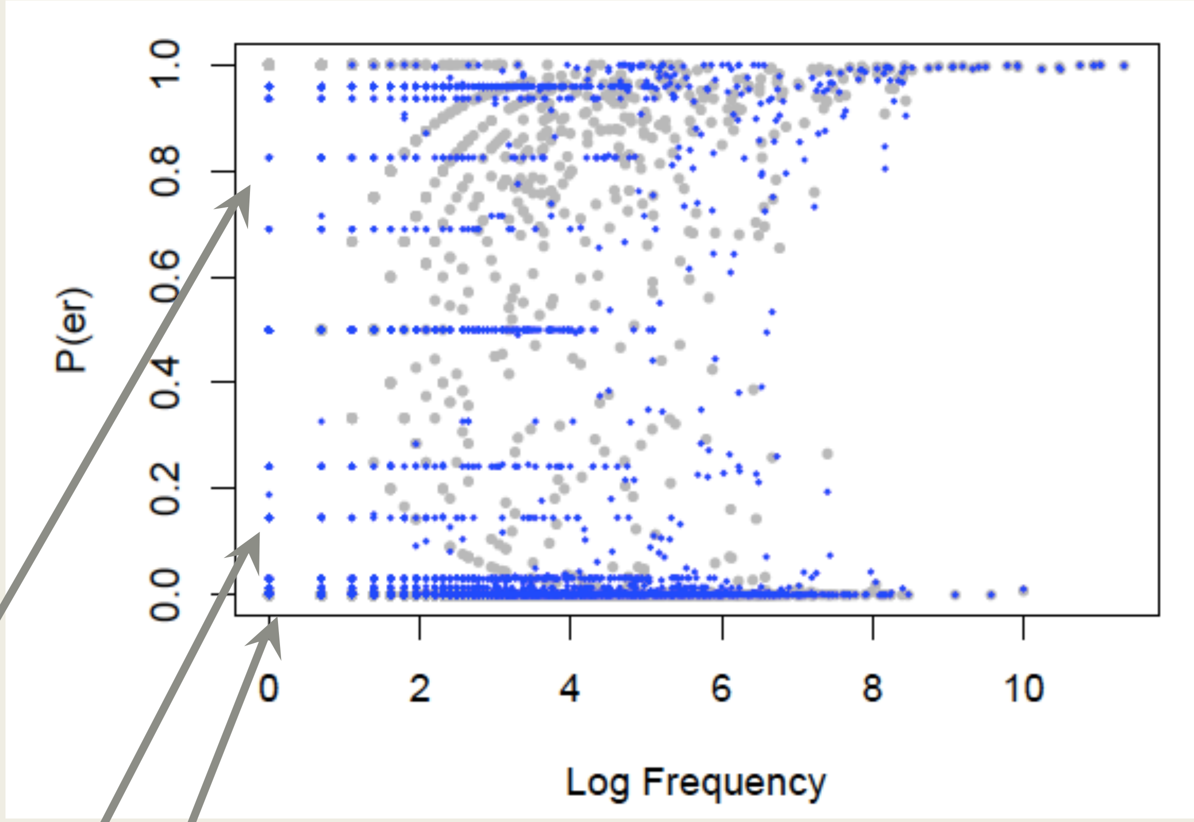
Constraints:

- 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)



1 syllable, -CC

2 syllables, *-r*

3+ syllables

COCA
(observed probabilities)

GLaPL
(predicted probabilities)

Higher frequency → More idiosyncratic

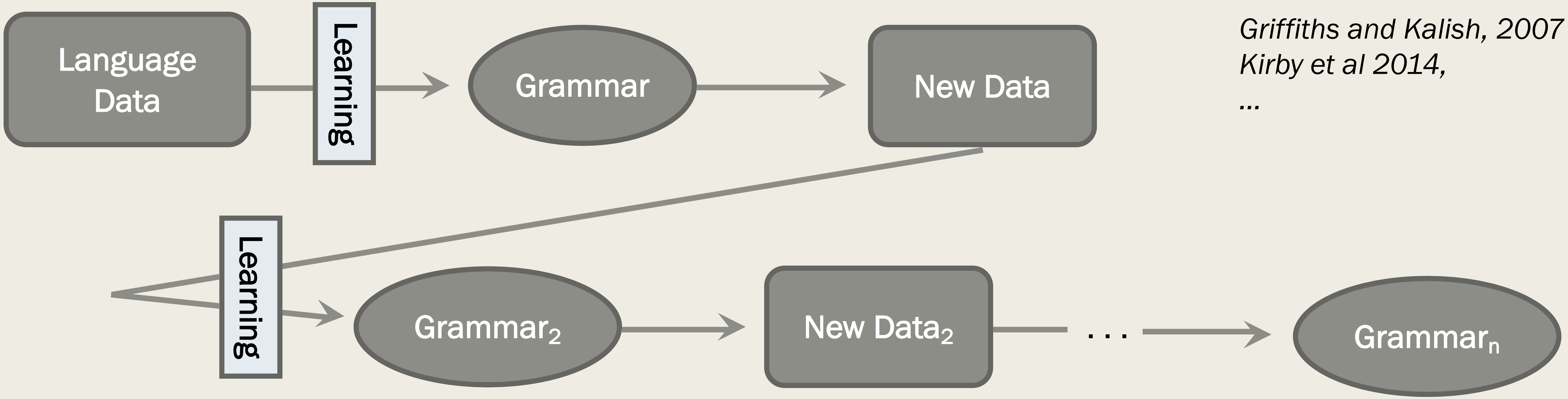
Lower frequency → Reliance on grammar

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***)

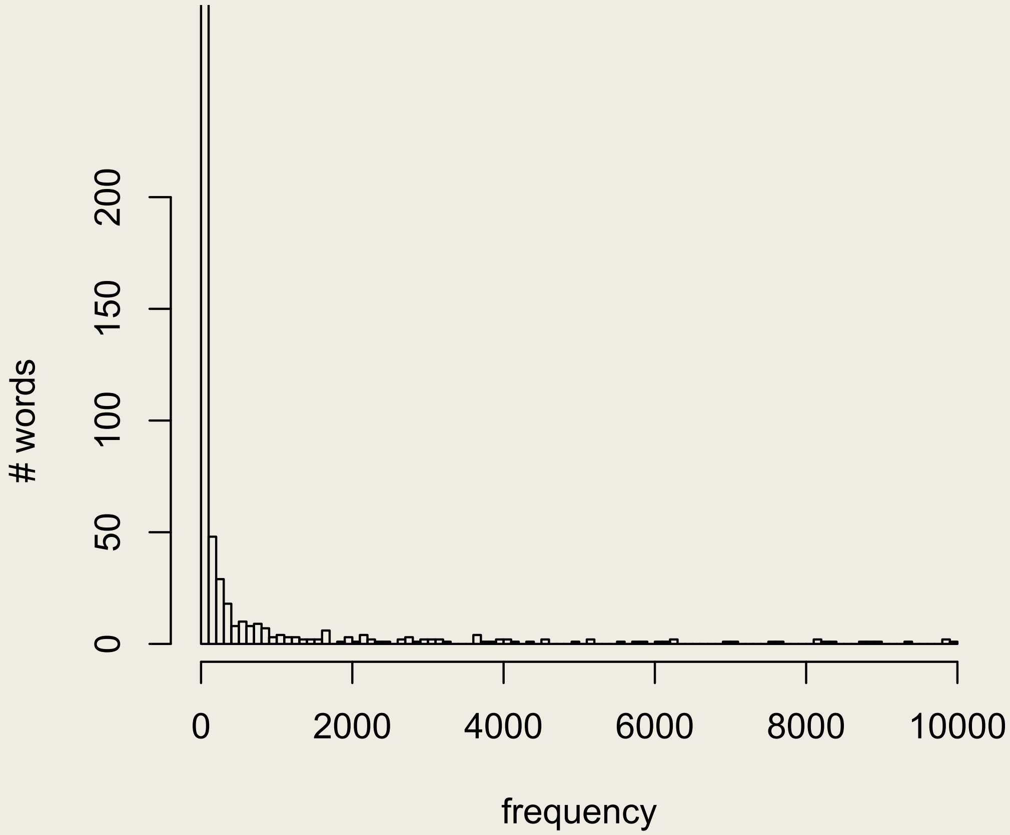


Griffiths and Kalish, 2007
Kirby et al 2014,
...

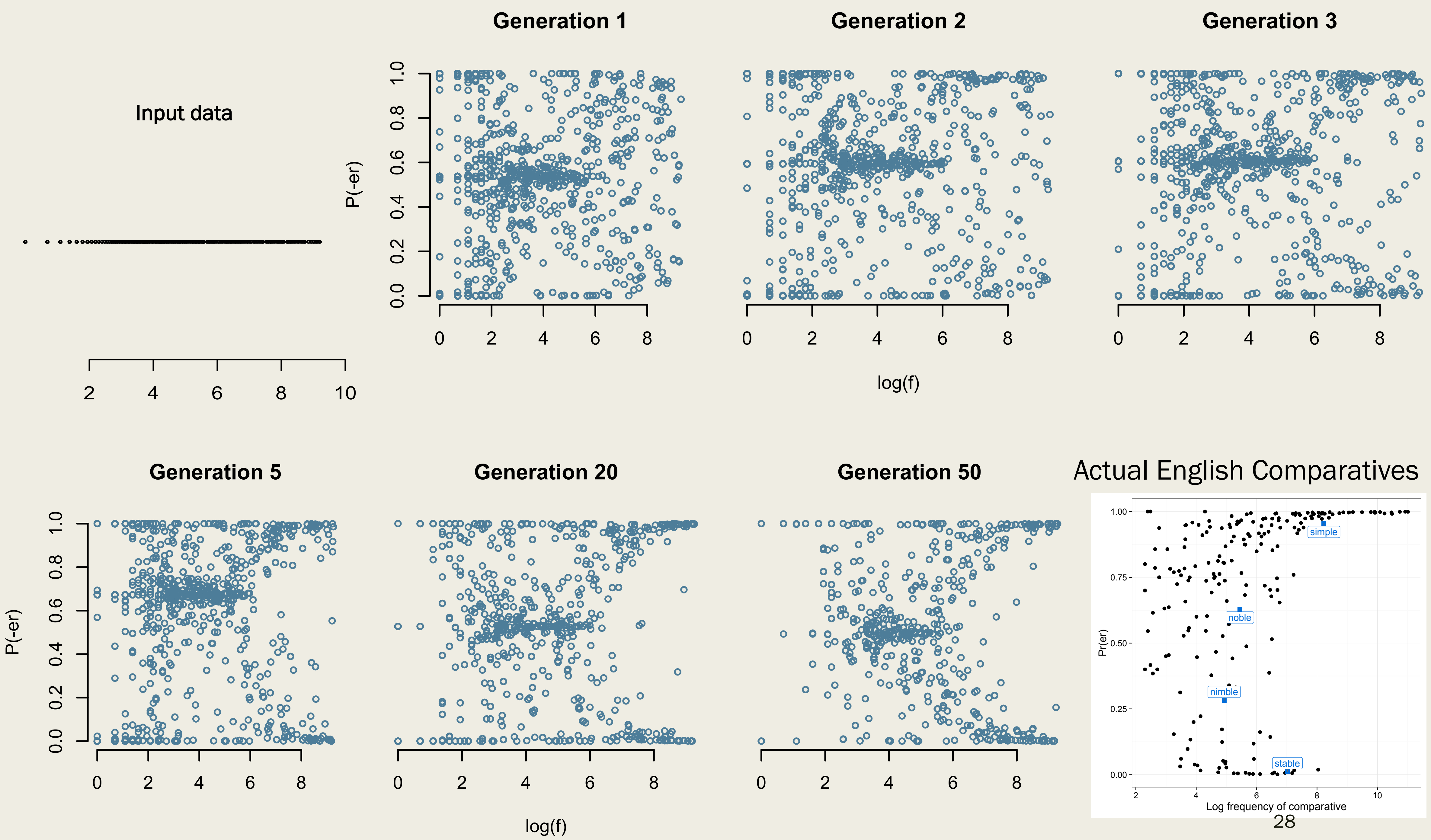
GLaPL: Exceptionality over generations

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

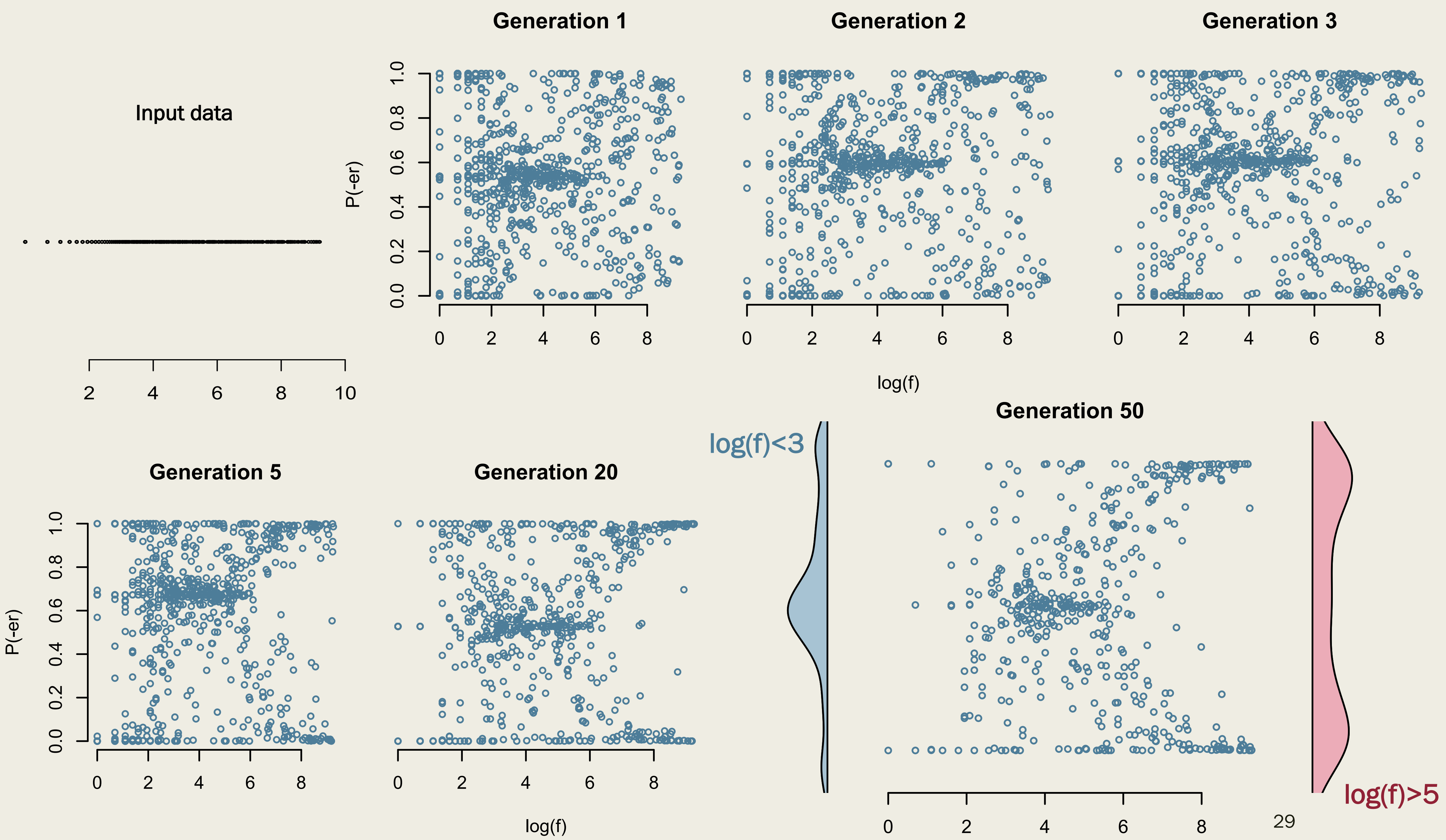
Two (relatively dumb) markedness constraints: **BE *more***, **BE *-er***



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**

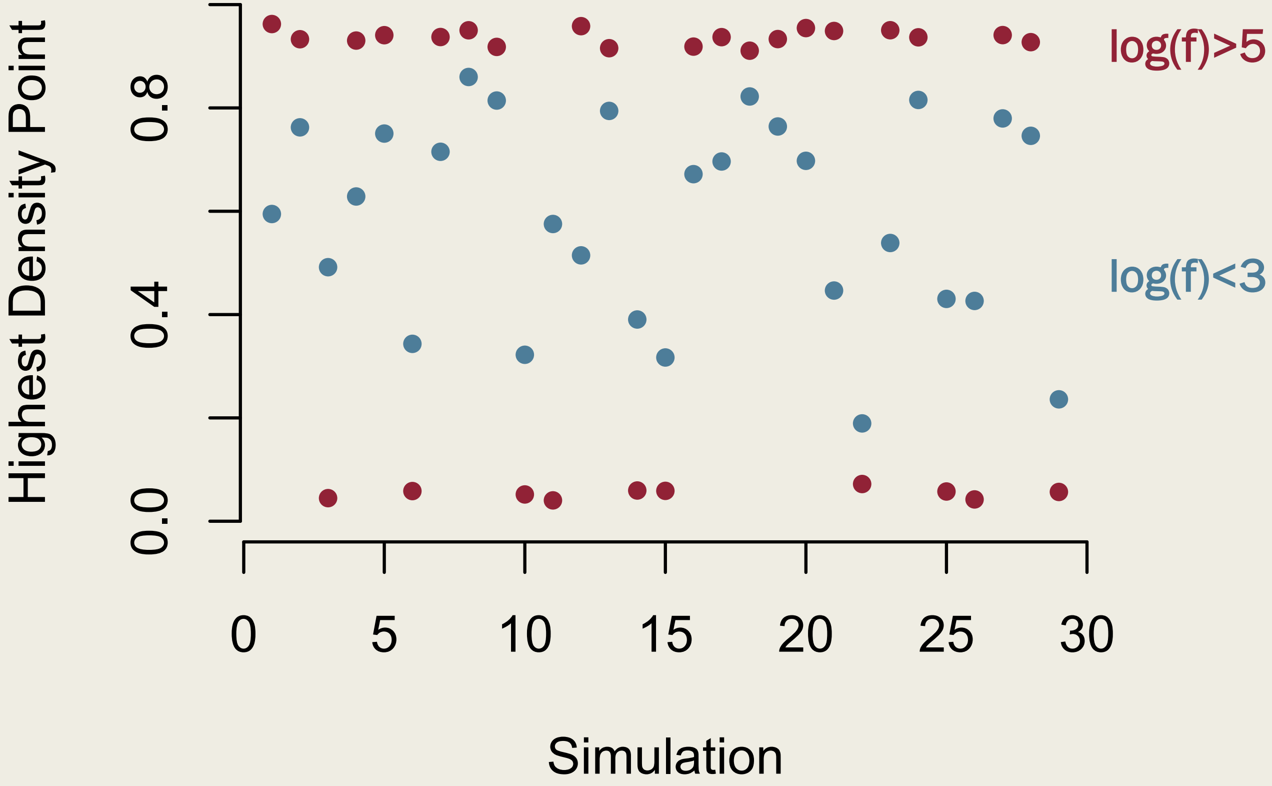


Input data
Generation 1
Generation 2
Generation 3
Generation 5
Generation 20
Generation 50
Actual English Comparatives
P(-er)
log(f)
Pr(er)
Log frequency of comparative
simple
noble
nimble
stable



Input data
Generation 1
Generation 2
Generation 3
Generation 5
Generation 20
Generation 50
P(-er)
log(f)
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 → more detailed representations
- Over time, detailed representations → 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
...

composed version

Pl

Pos1: -voice 8
Pos1: +sibilant 6.2
Pos1: +cont 3.1
...

	p	$\mathcal{H}$	*[-voi][+voi] 10	*Pos1 -voice 6.2*	*Pos2 +high 3.1*	...
People + Pl						
→ pipl	0.99	-0.2				
prsnz	0	-9.3		1	1	
prsns	0	-13.1	1		1	

stored version

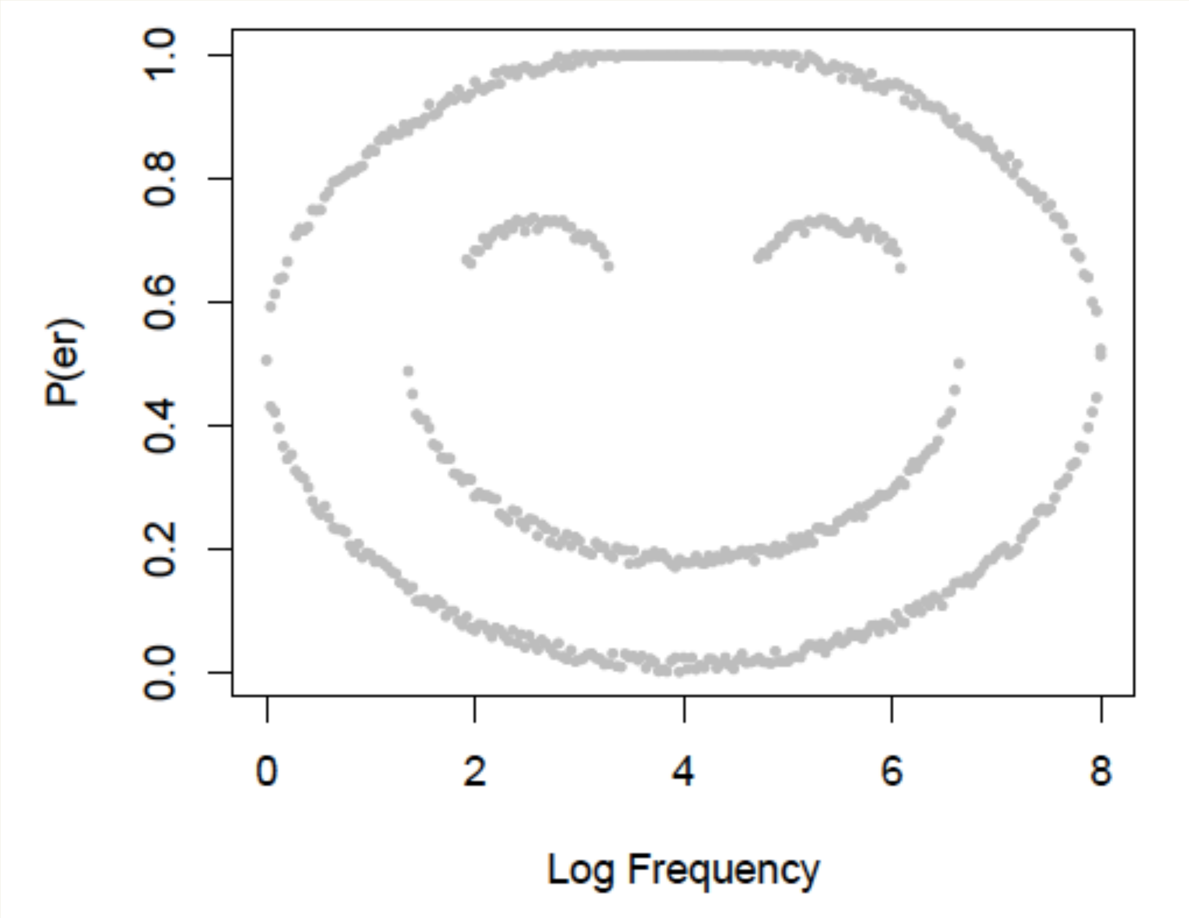
Person Pl *(people)*

Pos1: -voice 8
Pos1: +stop 6.2
Pos2: +high 7.9
Pos3: +stop 7.7
Pos4: -lateral 6
Initial stress 10
...

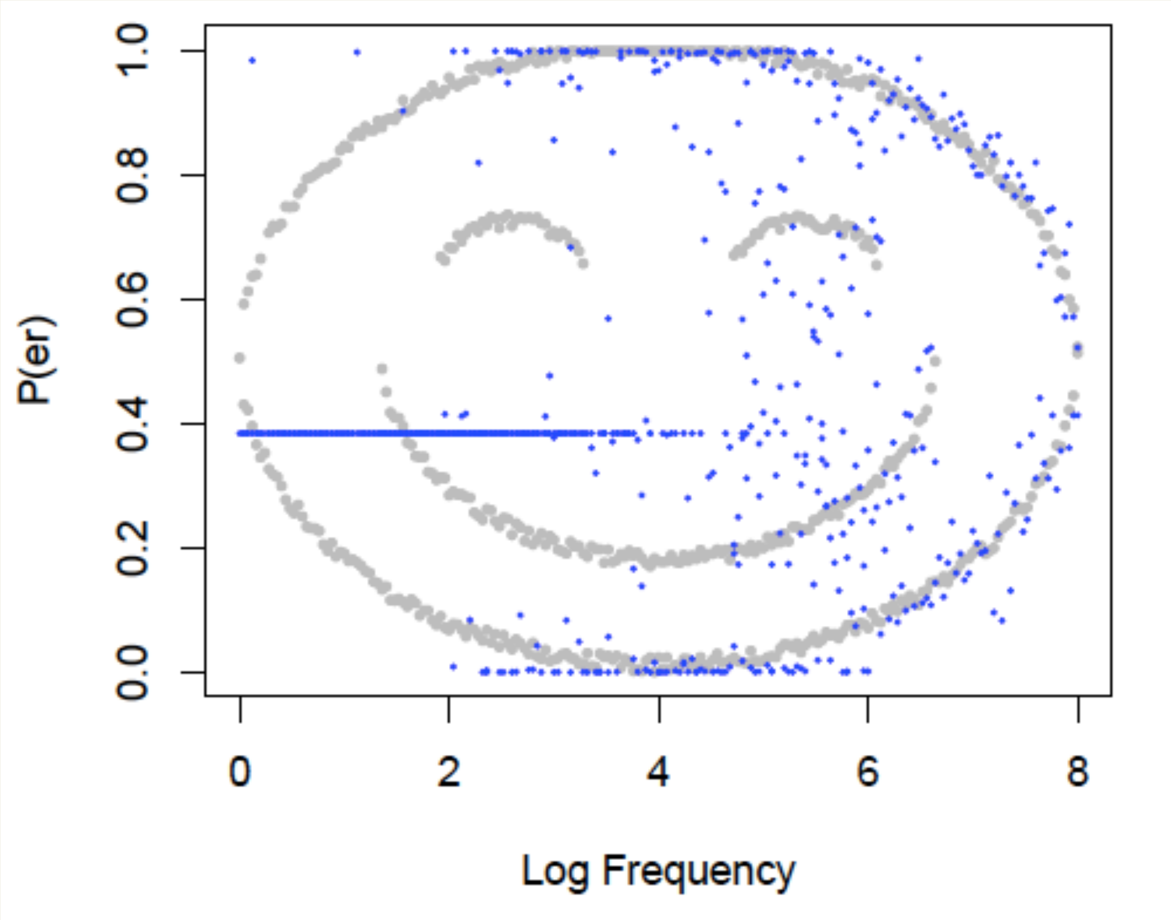
Choose between the stored version and the composed version however you want.

Markedness can overcome PFCs

	p	$\mathcal{H}$	*V́tV̆ 10	*Pos4 +stop 5*	*Pos4 +cor 10*	…	Pos1 +high 8	…
GREET + PROG								
→ gríɾɪŋ	0.99	-5		1				
grítɪŋ	0	-10	1					
grípɪŋ	0	-10			1			
gríɾəŋ	0	-13		1			1	

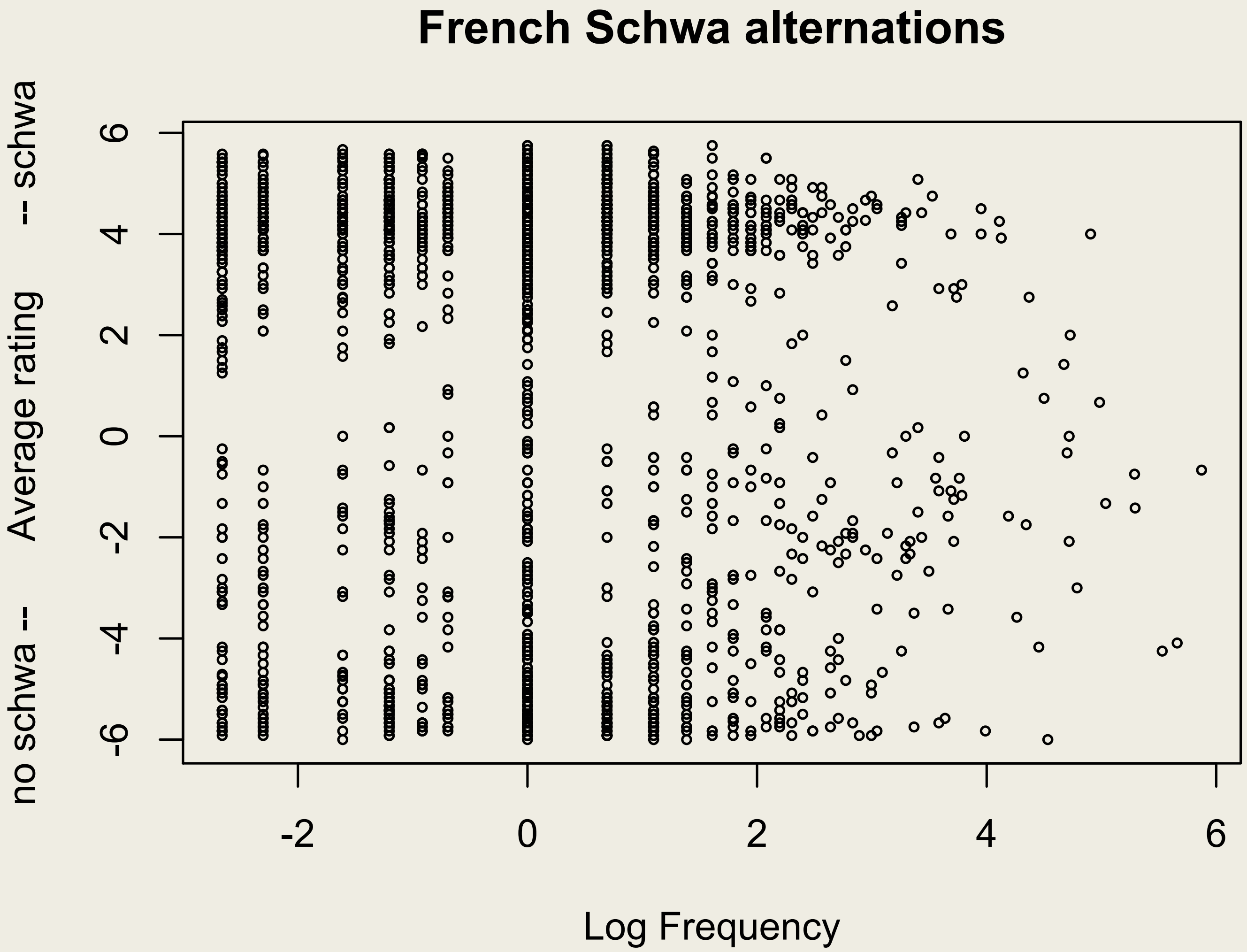


GLaPL trying to learn crazy data



GLaPL trying to learn crazy data

French schwa alternations



semaine ~ *smaine*
semetre ~ *smestre*

Data from *Racine, 2007*

12 Native speakers rated 2189 nouns with and without schwa