Logistic regression, the binomial construction, and a hierarchical regression model

Roger Levy

9.19: Computational Psycholinguistics

30 October 2023

Probing binomial ordering preferences

- In each pair, which phrase sounds more natural?

hit and run	*run and hit*
gold and silver	*silver and gold*
deer and trees	*trees and deer*
drink and food	*food and drink*
bacteria and candy	*candy and bacteria*
radio and television	*television and radio*
shares and stocks	*stocks and shares*
chanting and enchanting	*enchanting and chanting*
quails and felines	*felines and quails*

Ordering preferences in binomials

- Every occurring binomial is result of a *speaker's choice* about *binomial ordering*

(US Google Books ngram counts, 1960–2012; ~340B words)	Count	Count(Rev)
salt and pepper	568,951	32,082
cat and mouse	26,774	367
skirts and sweaters	1,763	1,707
bishops and seamstresses	<40	<40
few and unfavorable	<40	<40
principal and interest	120,034	50,032

- What is the representation of these ordering preferences?
- Are these preferences also *productive*?

An *n*-grams dataset from millions of books

(image credit Top of the List)

RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17]*†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

pages of 1208 books. The corpus contains 386,434,758 words from 1861; thus, the frequency is 5.5×10^{-5}. The use of "slavery" peaked during the Civil War (early 1860s) and then again during the civil rights movement (1955–1968) (Fig. 1B)

In contrast, we compare the frequency of "the Great War" to the frequencies of "World War I" and "World War II". References to "the Great War" peak between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as "World War I" (Fig. 1C).

These examples highlight two central factors that contribute to culturomic trends. Cultural change guides the concepts we discuss (such as "slavery"). Linguistic change, which, of course, has cultural roots, affects the words we use for those concepts ("the Great War" versus "World War I"). In this paper, we examine both linguistic changes, such as changes in the lexicon and grammar, and cul-

(Michel et al., 2011; the Google Books project)

Testing some more intuitions

boof and kaboof	*kaboof and boof*
glagy and gligy	*gligy and glagy*
swirp and swirr	*swirr and swirp*
smates and smats	*smats and smates*
rasby and dasby	*dasby and rasby*

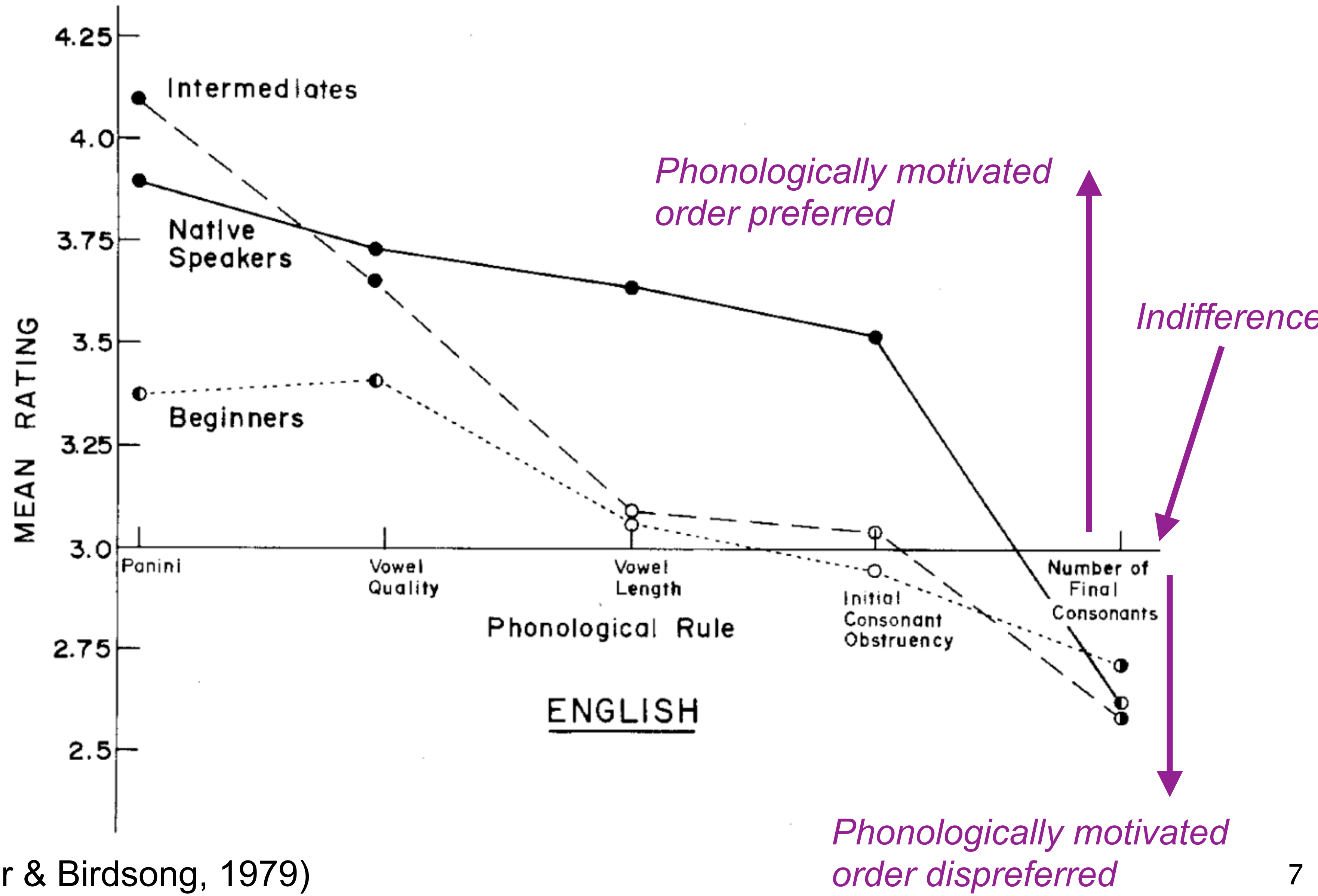
(Pinker & Birdsong, 1979)

Testing some more intuitions

fim	–	*fum*	*fum*	–	*fim*
begroast	*and*	*begroat*	*begroat*	*and*	*begroast*
spladilk	*or*	*dilk*	*dilk*	*or*	*spladilk*
waf	–	*paf*	*paf*	–	*waf*
frinning	*and*	*freening*	*freening*	*and*	*grinning*

(Pinker & Birdsong, 1979)

Ordering preferences for nonce words



(Pinker & Birdsong, 1979)

Previous work: *novel* binomials

- Pinker & Birdsong (1979) used *nonce-word* binomials to test phonological constraints in offline judgments:
  - ✓ Length (*boof and kaboof; *dadabig and dabig*)
  - ✓ Vowel Quality: high<low (*gligy and glagy*; **roppo and reppo*)
  - ✓ Vowel Length: long<short (*smats and smates*)
  - ✓ Initial Consonant: sonorant<obstruent (*haipo and daipo*)
  - × # Final Consonants (*skalk and skall; *flar and flard*)
- McDonald, Bock, and Kelly (1993) tested (mostly) *novel* binomials in offline judgments and production:
  - ✓ Animacy
  - × Length in production
  - ✓ Length in offline judgments

Ordering preferences: productive knowledge

What constraints predict relative preference for *X and Y* versus *Y and X* has been extensively investigated (Malkiel 1959, Bolinger 1962, Cooper & Ross 1975, Gustafsson 1976, Fenk-Oczlon 1989, Benor & Levy 2006)

- Iconic/scalar sequencing
  - what comes first happens first
  - *open and read* (a book); *hit and run* (auto); **hit and run* (baseball)

Attested but violates constraint

- Perceptual Markedness
  - animate, concrete, positive, … < inanimate, abstract, negative, …
  - *deer and trees; honest and stupid; *flora and fauna*
- Power
  - More culturally prioritized or “powerful” word comes first
  - *clergymen and parishioners; food and drinks; *clerks and postmasters*

*The **condiment rule** (Cooper & Ross 1975)*

Ordering preferences: productive knowledge

- Formal Markedness
  - Words with more general or broader meaning distributions come first
  - *sewing and quilting*; *changing and improving;*roses and flowers*
- No final stress
  - The final syllable of *Y* in *X and Y* must not be stressed
  - *abused and neglected; skirts and sweaters; *manufacture and install*
- Frequency
  - The more frequent word comes first
  - *bride and groom; smile and wink; *psychiatrists and patients*
- Length ("Panini's Law")
  - The shorter word comes first (we count in syllables)
  - *ask and answer; tense and irritable; *family and friends*

Formalizing ordering preferences

- Varieties of *probabilistic grammar* for forms F and meanings M:
  - Grammars over *forms*: $P(F)$ (word strings, syntax trees, …)
  - Grammar over *possible forms given a meaning to be expressed*: $P(F \mid M)$
  - Interpretive grammars of possible meanings given a form: $P(M \mid F)$

$$P(\text{"}X \text{ and } Y\text{"} \mid \{X, Y\})$$

e.g., $P(\text{"pepper and salt"} \mid \{\text{salt, pepper}\})$

A dataset of binomial expressions

Binomials are all over in naturalistic use→easy to sample:

ask	and	answer	right	and	good
knew	and	admired	sit-ups	and	push-ups
medicines	and	yeast	fits	and	starts
surprised	and	dubious	anxiously	and	eagerly
rank	and	file	congressional	and	presidential
thick	and	brown	toe	and	fronts
understand	and	share	startling	and	skillful
consider	and	rate	carefully	and	prudently
commoners	and	kings	WordPerfect	and	Lotus
always	and	everywhere	milk	and	honey
stained	and	waxed	improperly	and	unfairly
officially	and	publicly	business	and	government
tear	and	inflame	playbacks	and	study
By	and	large	cold	and	wet
linguistic	and	paralinguistic	softly	and	triumphantly
further	and	unnecessarily	register	and	vote
pie	and	bar	proposed	and	accepted
anger	and	anxiety	geographical	and	socio-economic
follow	and	understand	welcomed	and	approved
crime	and	sports	dwindling	and	diminishing
poetry	and	non-poetry	tough	and	dirty
immediately	and	directly	eighth	and	ninth

⋮

(Benor & Levy, 2006)

Probabilistic models of binomial ordering preferences

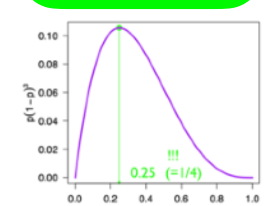
- One-constraint model, e.g.,

$$P([\textsc{short}]\ \textit{and}\ [\textsc{long}] \mid \{[\text{short}],[\text{long}]\}) = p$$

- In our dataset, 65% preference when conjuncts differ in number of syllables
  - We set the relative-frequency estimate of p to 0.65
  - Remember: this is the ***maximum likelihood estimate!***

abused and neglected	✔	*people and soils*	✘
bold and entertaining	✔	*surprised and dubious*	✔
coughed and chattered	✔	*sought and received*	✔
medicines and yeast	✘	*sharp and rapid*	✔

MLE



From earlier in the semester!

Multiple, cross-cutting constraints

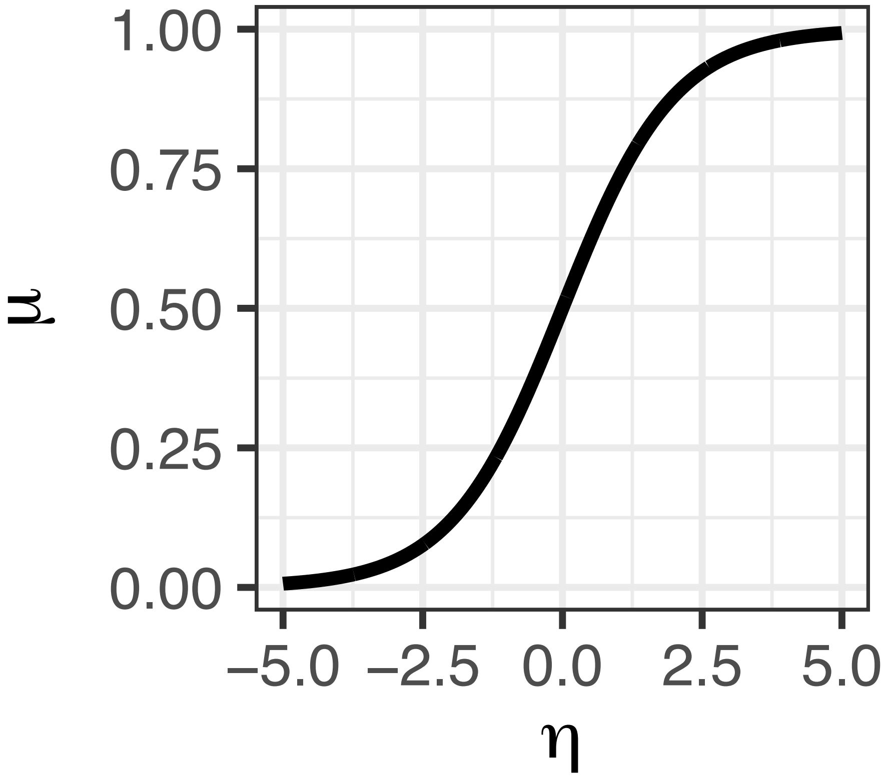
- When we have more constraints, we use ***logistic regression***

$$\underline{P(\text{“success”})} = \frac{e^{\eta}}{1+e^{\eta}}$$

a.k.a. **mean** μ

$$\underline{\eta} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N$$

a “goodness score”



Logistic (sigmoid) activation function

Fitting logistic regression via MLE

- With multiple model parameters, we get a likelihood *surface* on which we want to find the maximum
- 2-constraint example: word **length** and word **frequency**

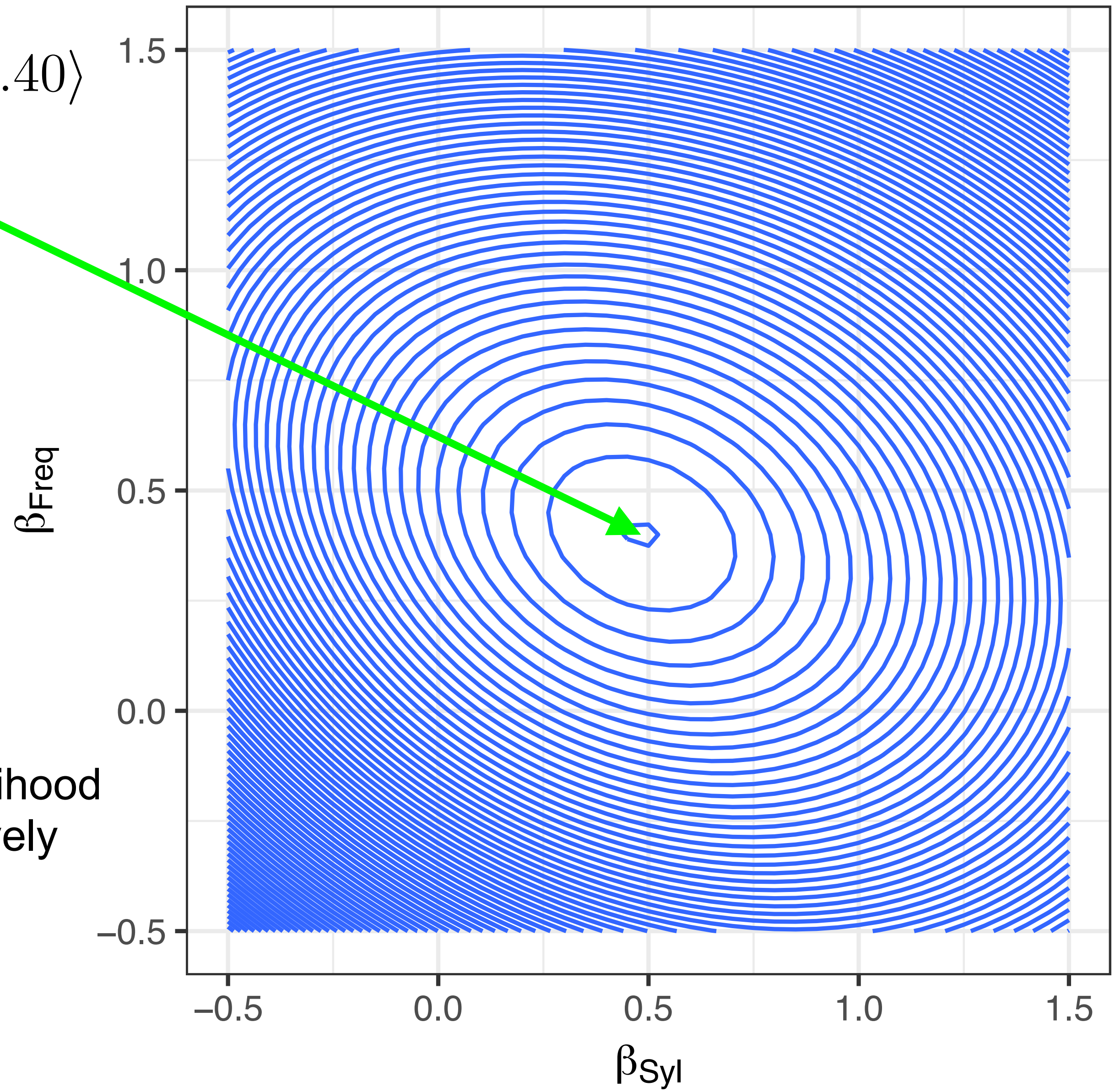
	Short<Long?	**Freq<Infreq?**
new and modern	✔	✔
correct and acute	***n/a***	✔
down and out	***n/a***	✘
cruel and unusual	✔	✘
anger and spite	✘	✔
crochet and knit	✘	✘

$$\eta = \beta_{\mathsf{Syl}} X_{Syl} + \beta_{\mathsf{Freq}} X_{Freq}$$

$$P(\text{A and B} | \{A, B\}) = \frac{e^{\eta}}{1 + e^{\eta}}$$

Maximum of the likelihood surface

$$\langle \widehat{\beta}_{Syl}, \widehat{\beta}_{Freq} \rangle = \langle 0.48, 0.40 \rangle$$



For logistic regression, likelihood surface is **convex** — relatively easy to find optimum

Model predictions from fitted parameters

Logistic Regression Model Structure

$$\eta = \beta_{\text{Syl}} X_{Syl} + \beta_{\text{Freq}} X_{Freq}$$

$$P(\text{A and B} | \{A, B\}) = \frac{e^\eta}{1 + e^\eta}$$

a.k.a. **mean** μ

Fitted model parameters

$$\langle \widehat{\beta}_{Syl}, \widehat{\beta}_{Freq} \rangle = \langle 0.48, 0.40 \rangle$$

Model predictions

	Short<Long	Freq<Infreq?
new and modern	✔	✔
correct and acute	n/a	✔
down and out	n/a	✘
cruel and unusual	✔	✘
anger and spite	✘	✔
crochet and knit	✘	✘

Multiple, cross-cutting constraints

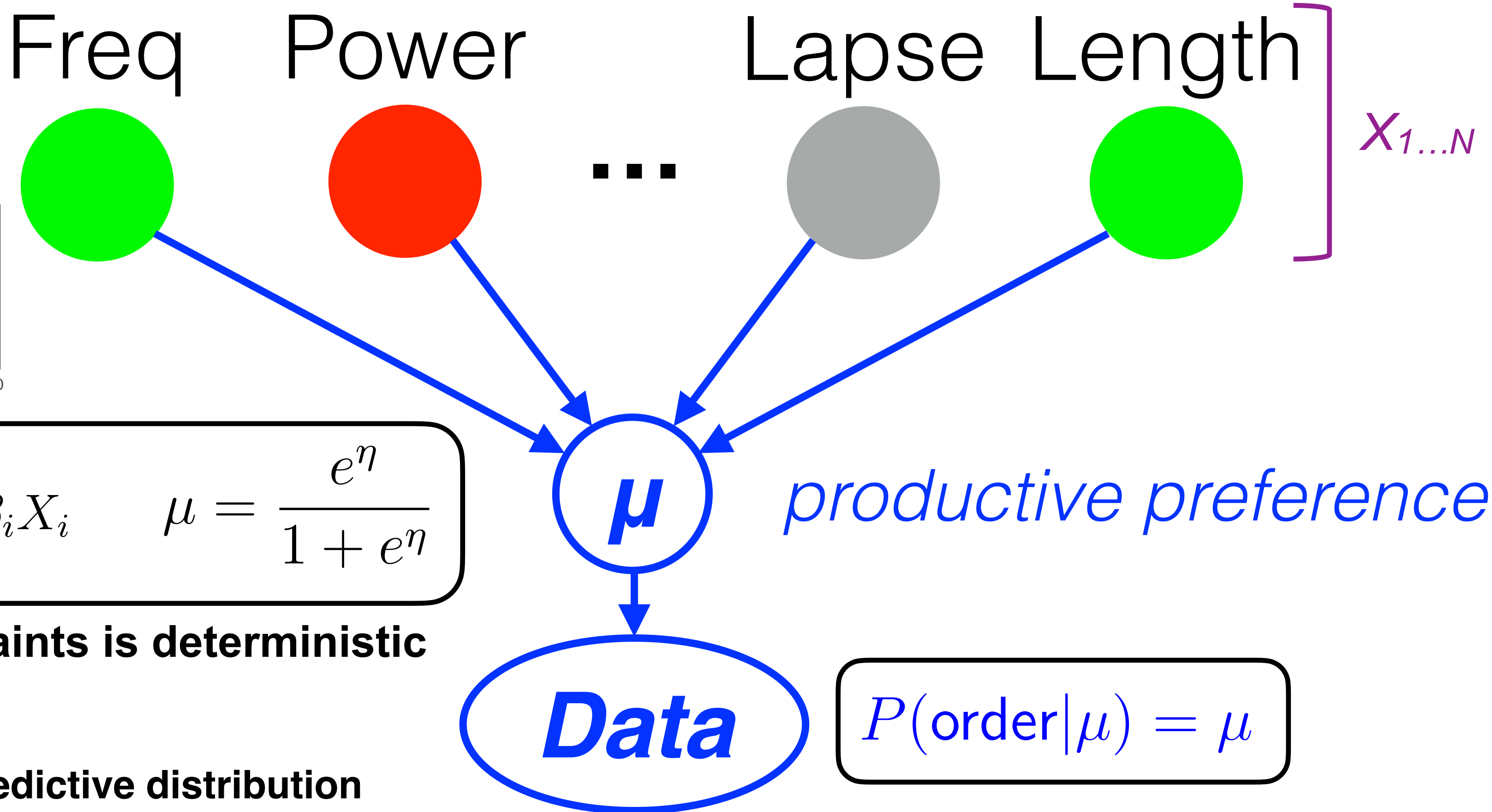
$\{X_i\}$

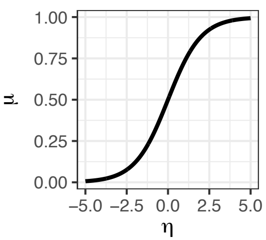
Constraint	Example	Strength
Iconic/scalar sequencing	*open and read*	**20**
Perceptual markedness	*deer and trees*	**1.7**
Formal markedness	*change and improve*	**1.4**
Power	*food and drink*	**1**
Avoid final stress	*confuse and disorient*	**0.5**
Short<Long	*cruel and unusual*	**0.4**
Frequent<Infrequent	*neatly and sweetly*	**0.3**

$\{\beta_i\}$

(from Morgan & Levy, 2016)

As a Bayes Net:





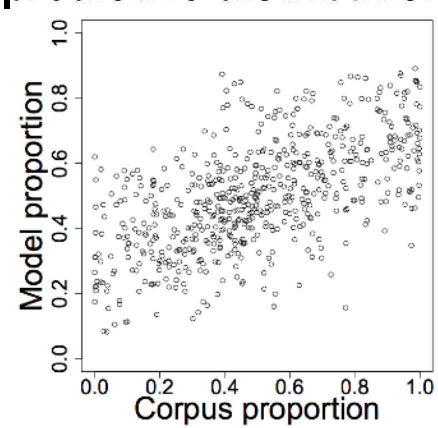
$$\eta = \sum_i \beta_i X_i \qquad \mu = \frac{e^\eta}{1+e^\eta}$$

μ|Constraints is deterministic

$$P(\text{order}|\mu) = \mu$$

Bernoulli (coin-flip) distribution

predictive distribution



Another source of knowledge

seamstresses and bishops
meat and potatoes — corpus prob | {meat, potatoes}≈0.95

OR

bishops and seamstresses
potatoes and meat — corpus prob | {meat, potatoes}≈0.05

?

You may prefer this because you're biased toward:

- culturally more powerful/central before less powerful/central
- short before long
- frequent before infrequent
- minimizing # consecutive unstressed syllables

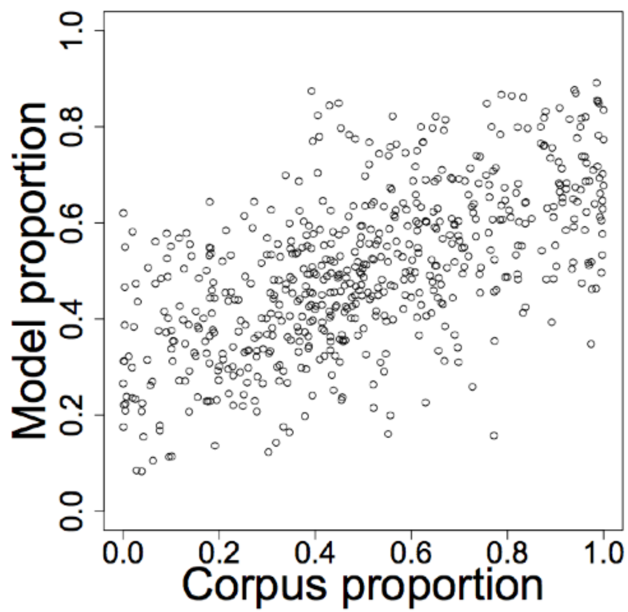
Productive knowledge

OR, you may prefer it before you've heard it far more often!

Direct experience

Productive knowledge and direct experience

- Our logistic regression model isn't perfectly predictive



- Part of this is that it fails to capture idiosyncrasy from direct experience
- A rational learner should…
  - …apply productive knowledge in novel expressions
  - …rely more on direct experience when it's plentiful

Binary forced-choice experiment

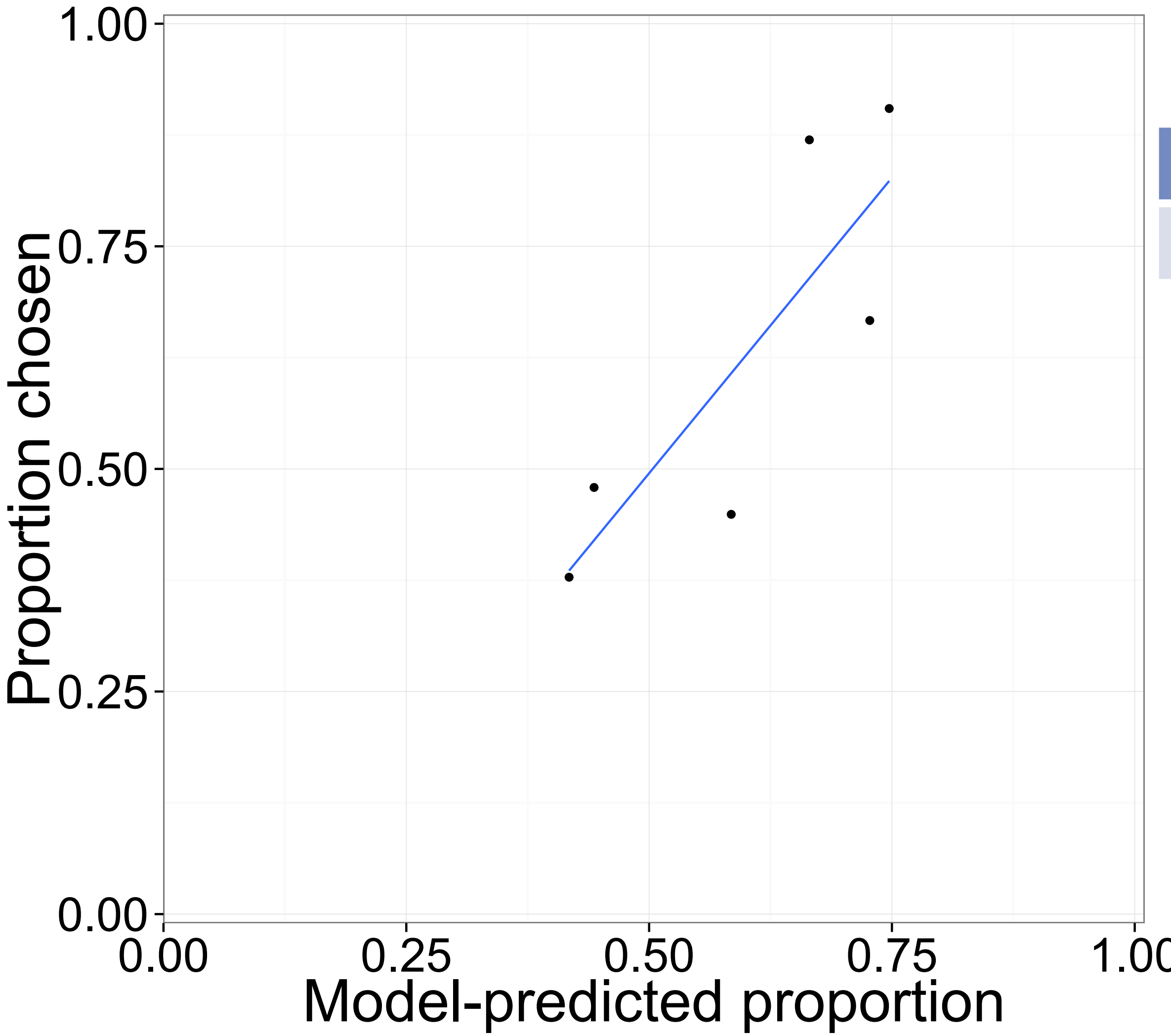
"Which sounds better?"

*There were many **bishops and seamstresses** in the small town where I grew up.*

*There were many **seamstresses and bishops** in the small town where I grew up.*

(Morgan & Levy, 2016, Cognition & unpublished)

Results: novel binomials



Predictor	Estimate	*p*
Abs know	6.18	0.003**

(Morgan & Levy, 2016, Cognition & unpublished)

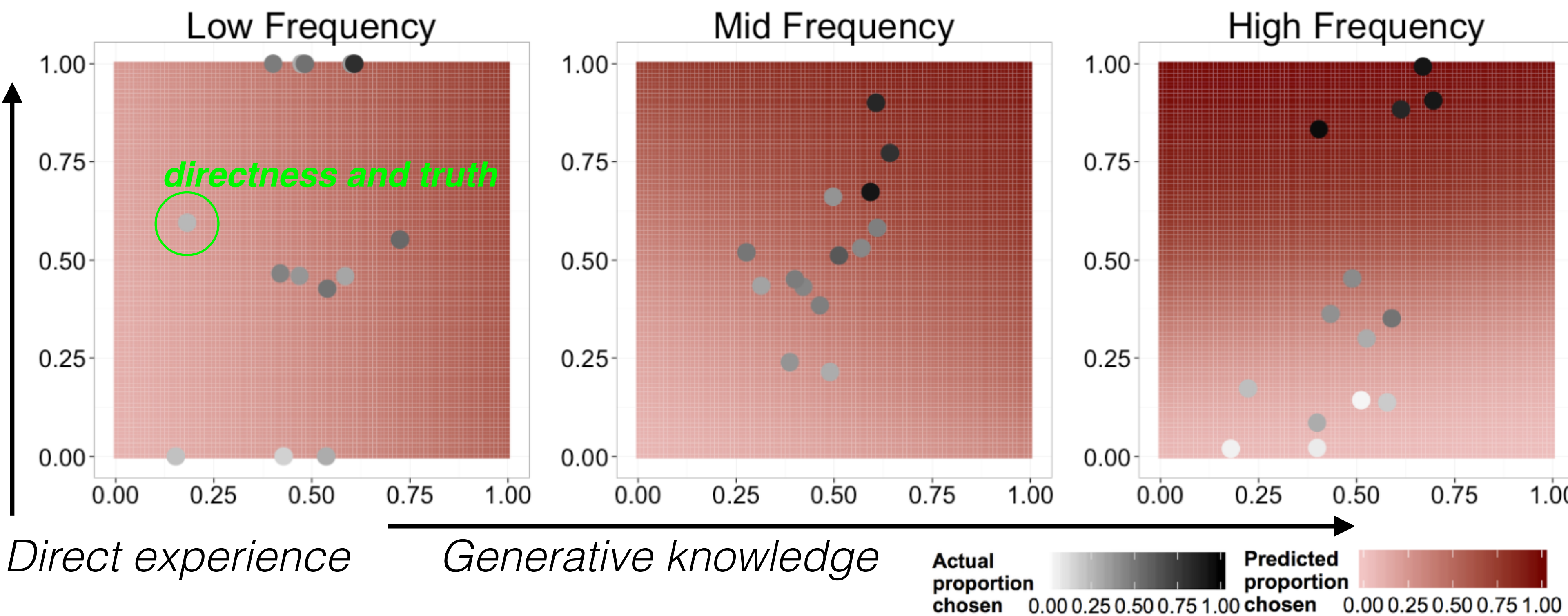
Results: attested binomials

=statistically significant (reliably non-zero)

Predictor	Estimate
Direct experience	*0.99**
Gen. knowledge	*2.36**

Predictor	Estimate
Direct experience	*3.32***
Gen. knowledge	*1.73*

Predictor	Estimate
Direct experience	*6.71****
Gen. knowledge	*-0.61*



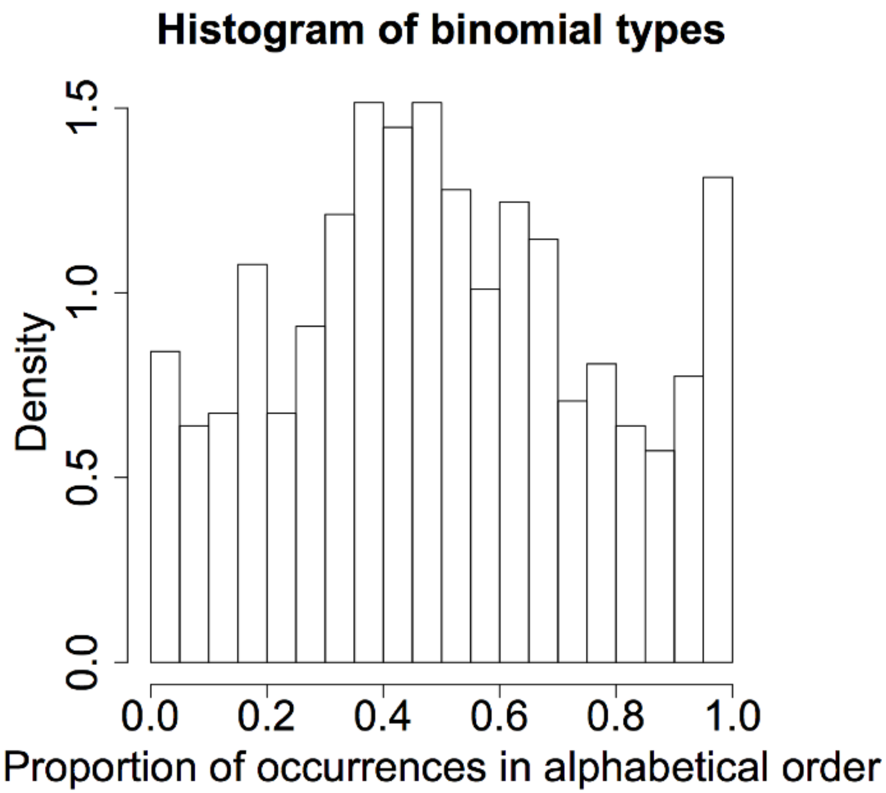
(Morgan & Levy, 2016, Cognition & unpublished)

The idiosyncratic and the general

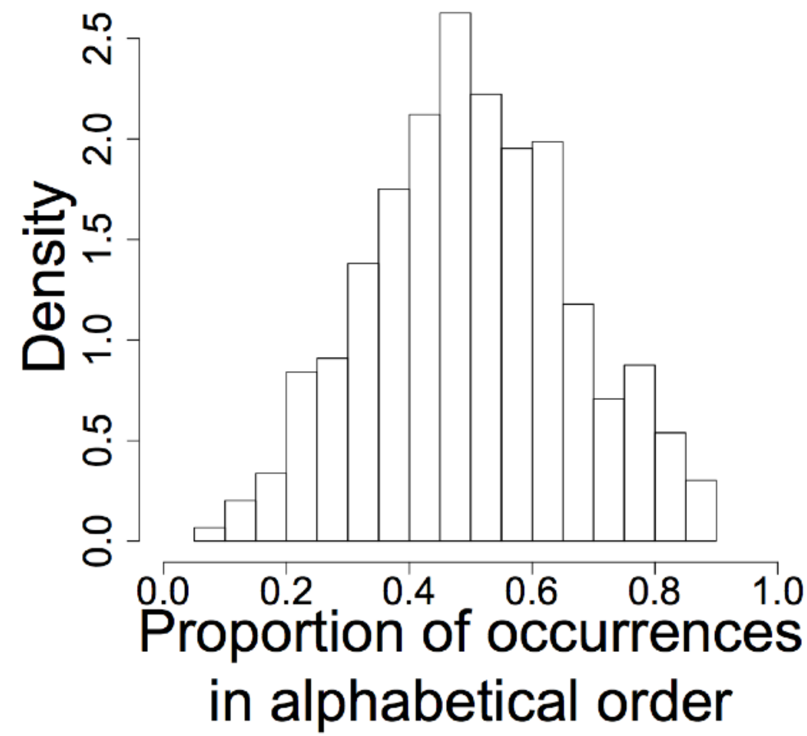
- We've seen evidence that binomial-specific ordering preferences have cognitive reality for speakers
- How dramatically do these preferences depart from the overall generative knowledge?
- How can we model both the generative knowledge and the idiosyncratic preferences simultaneously?

Distribution of ordering preference

Reality



Our model



Ordering preferences depart dramatically from generative knowledge!

(Morgan & Levy, 2015)

Modeling idiosyncrasy

- Here was logistic regression:

$$P(\text{“success”}) = \frac{e^{\eta}}{1+e^{\eta}}$$

$$\eta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N$$

- We revise it to include a ***beta-binomial*** *component*

$$P(\text{“success”}) = p$$

$$p \sim \text{Beta}\left(\frac{e^{\eta}}{1+e^{\eta}}, \nu\right)$$

$$\eta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N$$

(Morgan & Levy, 2015)

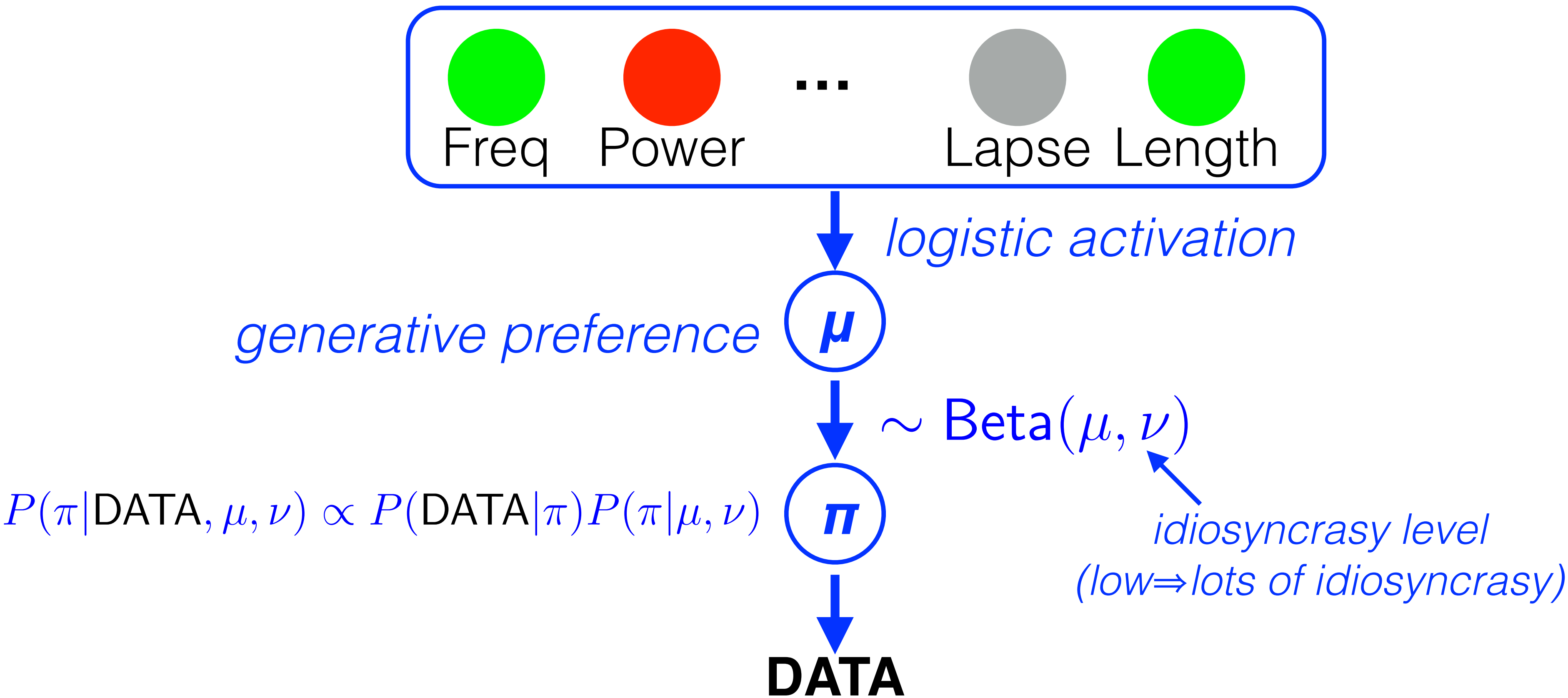
Frequency-sensitivity of binomial idiosyncrasy

Overall unordered frequency

$$\nu = exp(\alpha + \beta \cdot \log(M_n))$$

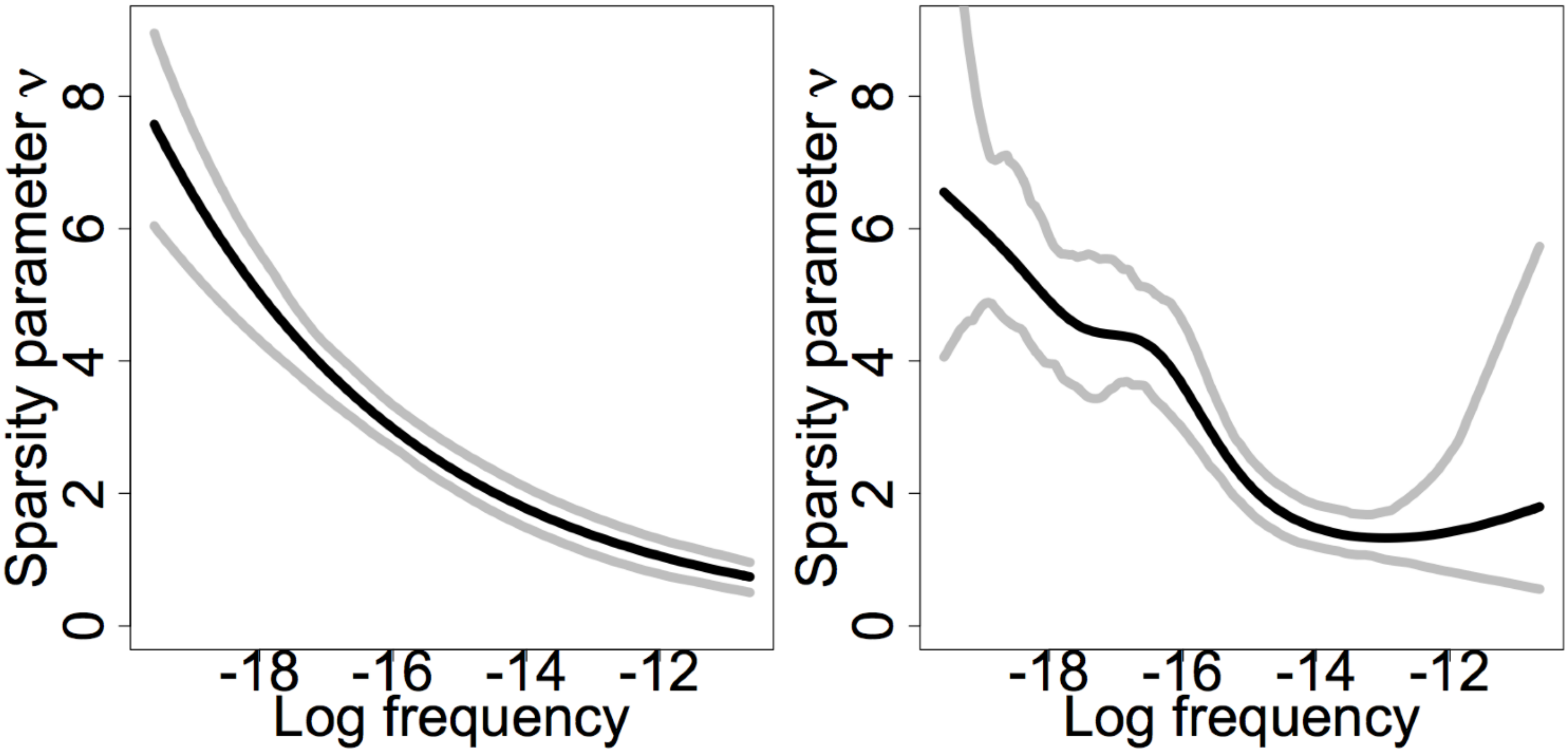
(Morgan & Levy, 2015)

Our complete model



(Morgan & Levy, 2015)

Results: frequency sensitivity of *v*

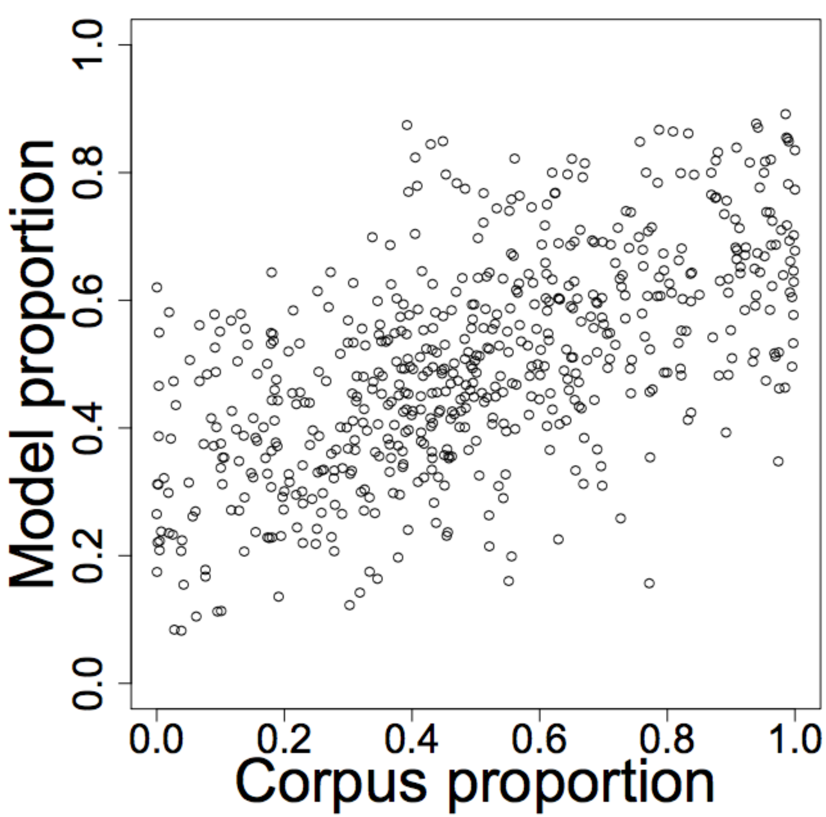


We call this ***frequency-sensitive regularization***
of binomial ordering preference

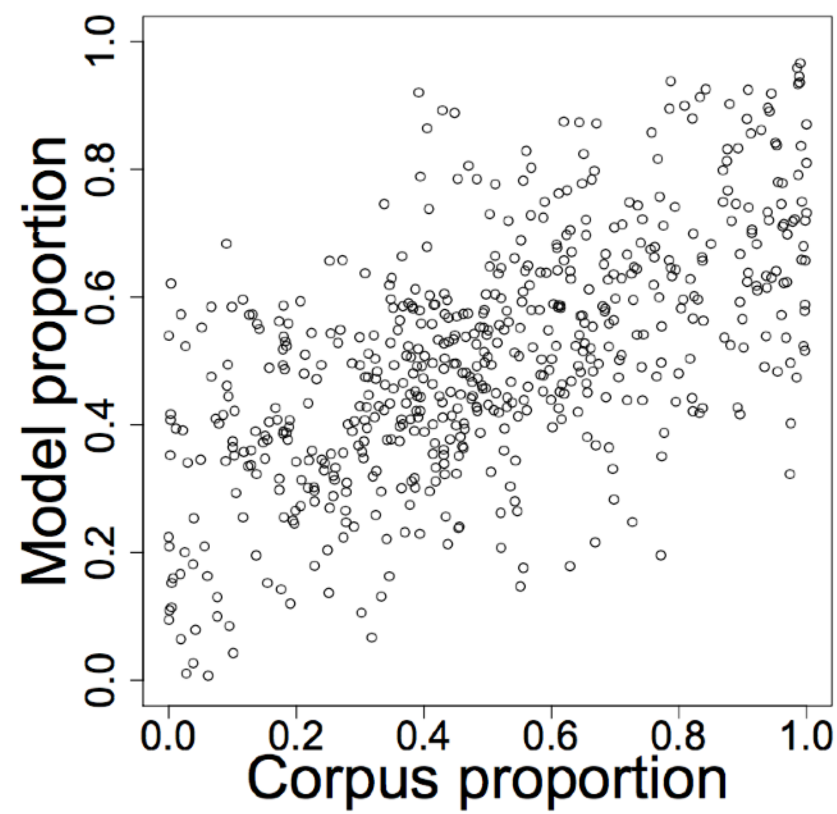
(Morgan & Levy, 2015)

Results: “best-guess” of preferences

Our OLD model



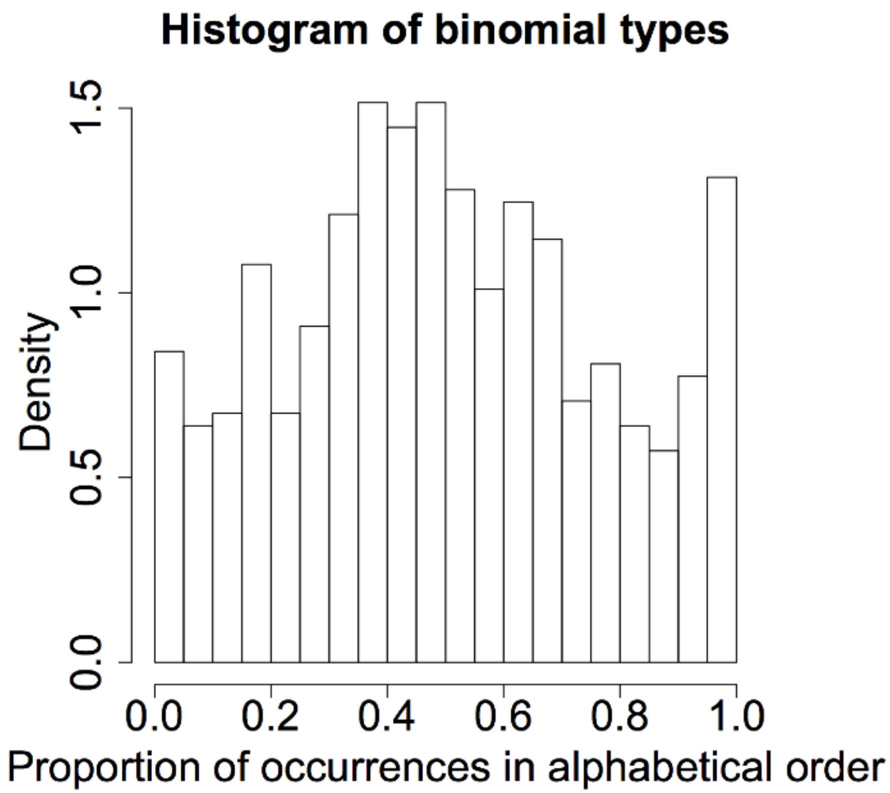
Our NEW model



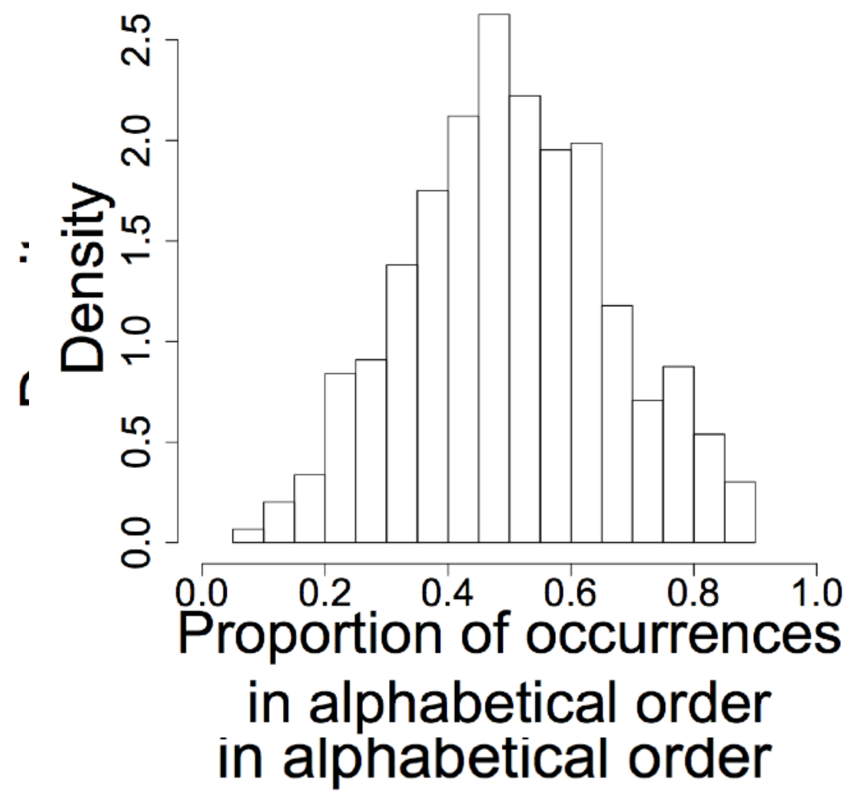
(Morgan & Levy, 2015)

Results: distribution of binomial prefs.

Reality



Our NEW model



(Morgan & Levy, 2015)

Summary for today

- In language we must often model **multiple**, **overlapping**, **defeasible** constraints that drive preferences
  - One example: linear **ordering preferences**
    - e.g., linear ordering preferences in the **binomial construction**
- We can do this with **logistic regression**
- Viewed as a **Bayes Net**, logistic regression imposes a **parametric form** on P(outcome|$X_{1...m}$)
- Logistic regression is extendable with a **hierarchical** component to handle item-specific idiosyncrasies
  - One version of this: **beta-binomial regression**

References

Agresti, A. (2002). Categorical data analysis. John Wiley & Sons.

Agresti, A. (2007). An introduction to categorical data analysis (Vol. 135). New York: Wiley.

Benor, S., & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. Language, 82(2), 233-278.

Cooper, W. E., & Ross, J. R. (1975). World order. Papers from the parasession on functionalism, 63-111.

McDonald, J. L., Bock, K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, *25*(2), 188-230.

Morgan, E., & Levy, R. (2015). Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *CogSci*.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384-402.

Pinker, S., & Birdsong, D. (1979). Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 497-508.