

# What's in a word?

## Using computational modeling to study phonotactic learning

Connor Mayer

UCI Department of Language Science

CTBS Seminar Series  
May 15, 2025



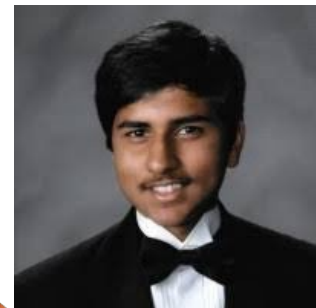
# About me



- Computational [phonologist and phonetician]
- Training in linguistics and computer science from the University of British Columbia
- Worked as a software developer on big budget video games for about 4 years
- Did my PhD in linguistics at UCLA
- Assistant prof in UCI Language Science
- I study speech!
  - Phonotactic learning
  - Speech biomechanics and motor control
  - Variability in phonological patterns
  - The Uyghur language

# Collaborators

This work is part of a larger NSF-funded project  
(#2214017) with Megha Sundara (UCLA)



# Goals of this talk

1. I want to teach you something about phonotactics!
2. I want to illustrate a general approach to theory comparison that takes computational models seriously as formal instantiations of linguistic theory

# Roadmap

1. Background on phonotactics
2. Study 1: Theory comparison using phonotactic models
3. Study 2: Infant acquisition of phonotactics
4. Discussion and take-aways

# Roadmap

1. Background on phonotactics
2. Study 1: Theory comparison using phonotactic models
3. Study 2: Infant acquisition of phonotactics
4. Discussion and take-aways

# Phonology

Phonology studies how sounds pattern within and across languages

Phonologists treat sound systems as formal symbolic systems

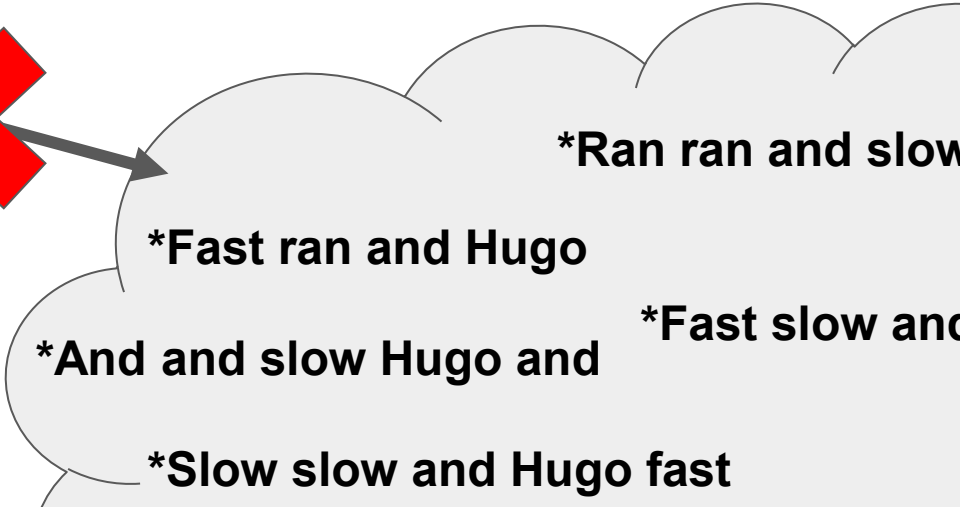
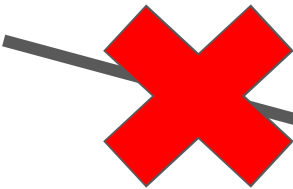
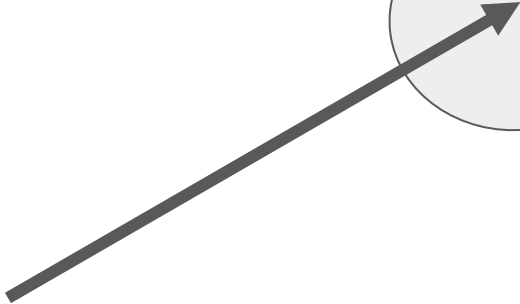
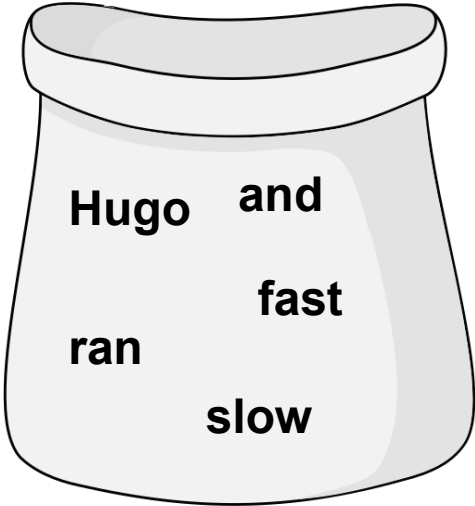
- Languages have a **finite set of sounds** from which words are formed

‘bash’ [bæʃ]

‘cache’ [kæʃ]

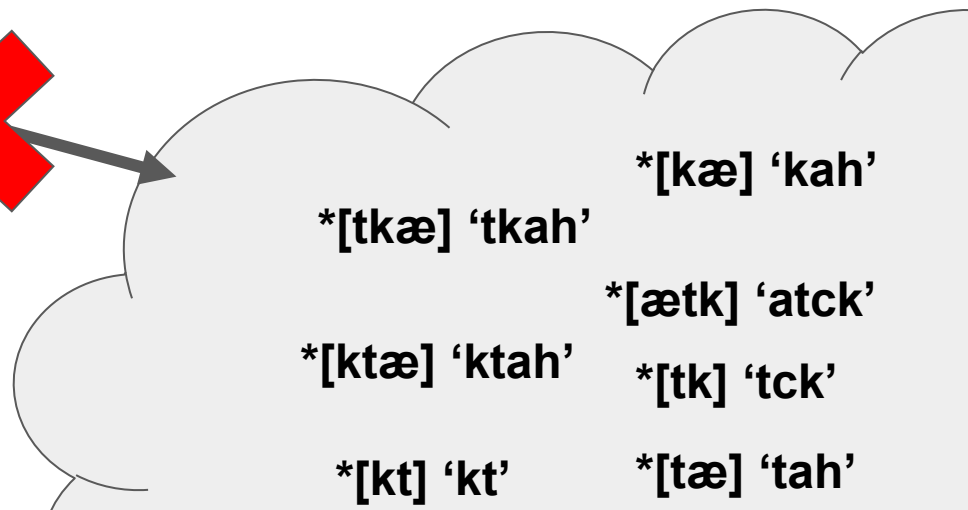
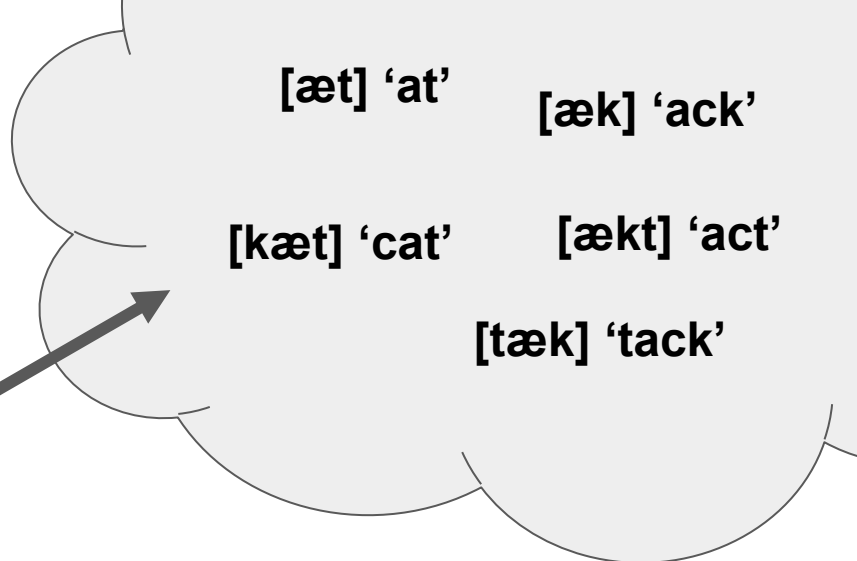
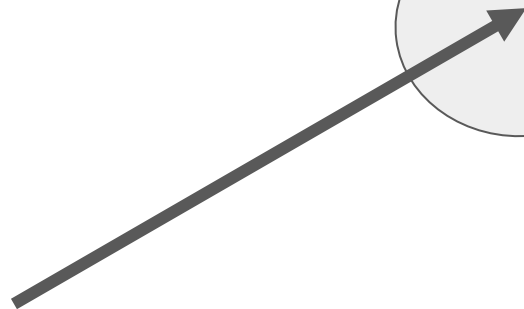
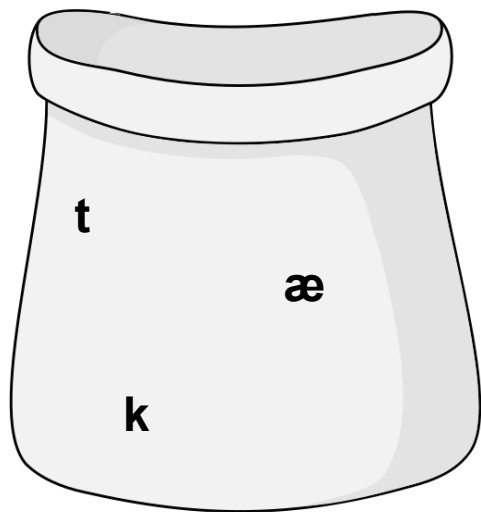
We’ll focus on a subdomain of phonology called **phonotactics**

Infinite use of finite means





# Infinite use of finite sounds



# Phonotactics

Restrictions on how sounds can be sequenced into words

This is (mostly) learned and language-specific:

- 'steek' would be a fine English word, but not a good Spanish word
- 'kwakwakuhwakw' is a fine Kwak'wala word, but not a likely English word

Speakers have implicit knowledge of the phonotactic properties of their language

# Probing phonotactic knowledge

A typical source of data is acceptability judgments

- “On a scale of 1-7, how likely is ‘steek’ to be an English word?”
- “Would ‘steek’ be a better English word than ‘kwakwakuhwakw’?”
- “Could ‘steek’ be an English word?”

These judgments consistently display *gradience* (Chomsky and Halle 1965, 1968, Coleman and Pierrehumbert 1997, Scholes 1966, Bailey and Hahn 2001, Hayes and Wilson 2008, Daland et al. 2011, a.o.)

What do we mean by gradient?

**poik**

**lvag**

**kip**

What do we mean by gradience?

**lvag ≪ poik ≪ kip**



# Roadmap

1. Background on phonotactics
2. Study 1: Theory comparison using phonotactic models
3. Study 2: Infant acquisition of phonotactics
4. Discussion and take-aways

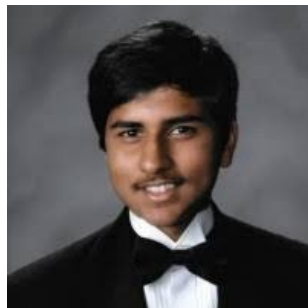
# Study 1: Theory comparison using phonotactic models

Mayer, Kondur & Sundara (accepted). The UCI Phonotactic Calculator: An online tool for computing phonotactic metrics. *Behavior Research Methods*.

Mayer, Wagner & Sundara (in prep). *Comparing segmental phonotactic models*.



Megha Sundara



Arya Kondur



Austin Wagner



# Learning phonotactic generalizations

**Broad question:** How is phonotactic knowledge operationalized?

- What dependencies and frequencies in the lexicon are we sensitive to?

We'll compare two computational models of phonotactic probability

- **Theoretical purpose:** what does this tell us about linguistic theory?
- **Practical purpose:** which tool is the most appropriate?

# Modeling phonotactic knowledge

**Goal**: we want a computational model that tracks with human behavior

All the models we consider treat phonotactics as probabilistic

$$P(w = x_1 \dots x_n)$$

**Output**: How probable is a word **w** composed of the segments **x<sub>1</sub>...x<sub>n</sub>**?

**Linking hypothesis**: Phonotactic probability correlates with acceptability ratings

# Other applications of phonotactic models

Phonotactic probability is relevant in many speech domains:

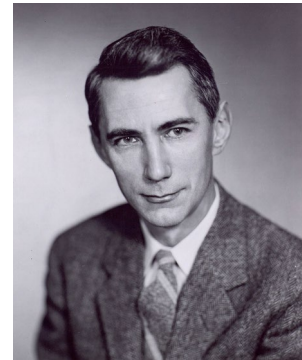
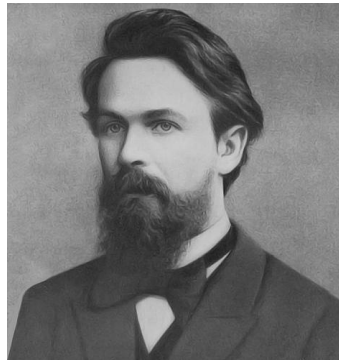
- **Speech perception** (e.g. Norris & McQueen 2008, Steffman & Sundara 2023)
- **Speech production** (e.g. Edwards et al. 2004)
- **Word segmentation and learning** (e.g. Mattys et al. 1999, Vitevitch and Luce 1999)
- **Speech errors** (e.g. Taylor & Houghton 2005, Goldrick & Larson 2008)
- **Sentence formation** (e.g. Hayes and Breiss 2020)

We often use phonotactic probabilities to model phenomena in these domains.

# Two prominent n-gram models

We'll compare two widely used n-gram models of phonotactics based on their ability to predict acceptability judgments

## 1. **Standard n-grams** (Markov 1913, Shannon 1948)



## 2. **Vitevitch & Luce's Phonotactic Probability Calculator** (Vitevitch and Luce 2004)



# Models as instantiations of theory

V&L describe their calculator as “relatively neutral with regard to linguistic theory”

Hayes (2012) claims that linguists would find it “extremely controversial”

- The standard n-gram model is less controversial from this perspective

These models are **formal instantiations of different theoretical assumptions**

- Our main goal is to assess the validity of these assumptions

# Why these models?

n-gram models are inadequate models of phonotactics, but still worth considering

Often ‘good enough’ for a quick and dirty quantification of phonotactics

- Bigram model on English acceptability judgment data  **$r = 0.877$**

(Daland et al. 2011, Dai, Mayer and Futrell 2023)

n-gram models are still widely used in research contexts

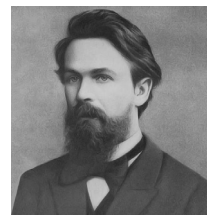
- Vitevitch and Luce (2004) has ~670 citations, ~160 from the last 4 years

## A note on historical precedence



Hayes, B. (2012). The role of computational modeling in the study of sound structure. Talk given at the 2012 Conference on Laboratory Phonology.

# The standard unigram model



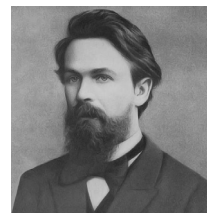
$$P_1(w = x_1 \dots x_n) \approx \prod_{i=1}^n P(x_i)$$



$$P_1(stik) = P(s) P(t) P(i) P(k)$$



# The standard bigram model

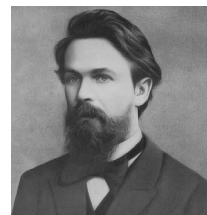


$$P_2(w = x_1 \dots x_n) \approx \prod_{i=2}^n P(x_i | x_{i-1})$$

#	s	t	i	k	#
---	---	---	---	---	---

$$P_2(\#stik\#) = P(s|\#) P(t|s) P(i|t) P(k|i) P(\#|k)$$

# Estimating probabilities from data



Add-one smoothed conditional probabilities from corpus counts

$$P(x) = \frac{C(x) + 1}{N + S}$$

**$C$** : Count function

**$N$** : Number of sound tokens

**$S$** : Number of sound types

$$P(y|x) = \frac{C(xy) + 1}{C(x) + S}$$

# The V&L unigram model



$$P_1(w = x_1 \dots x_n) \approx 1 + \sum_{i=1}^n P(w_i = x_i)$$

1	2	3	4
s	t	i	k

Below the table, there are four blue horizontal bars, each aligned under one of the letters s, t, i, and k.

$$P_1(stik) = P(w_1 = s) + P(w_2 = t) + P(w_3 = i) + P(w_4 = k) + 1$$

# The V&L bigram model



$$P_1(w = x_1 \dots x_n) \approx 1 + \sum_{i=2}^n P(w_{i-1} = x_{i-1}, w_i = x_i)$$

	1	2	3	4	
	s	t	i	k	

$$P_2(\#stik\#) = P(w_1 = s, w_2 = t) + P(w_2 = t, w_3 = i) + P(w_3 = i, w_4 = k) + 1$$

# Estimating probabilities from data in V&L



Token-weighted, joint probabilities from corpus counts

$$P_1(w_i = x) = \frac{C(w_i = x)}{C(w_i)}$$

$C$  is the token-weighted count

$$P_2(y|x) = \frac{C(w_{i-1} = x, w_i = y)}{C(w_{i-1}w_i)}$$

# Type-weighting vs. Token-weighting

Standard n-gram model is **type-weighted**

- Each occurrence of unigram/bigram contributes a count of 1

V&L model is **token-weighted**

- Unigrams/bigrams that occur in more frequent words count for more
- Count is equal to log frequency of word in corpus

# Type-weighting vs. Token-weighting

Corpus:

$$\begin{Bmatrix} \text{kæt: 1000} \\ \text{tæk: 50} \end{Bmatrix}$$

Type weighting:

$$C(\text{æ}) = 1 + 1 = 2$$

Token weighting:

$$C(\text{æ}) = \ln(1000) + \ln(50) = 10.82$$

# Summary of major model differences

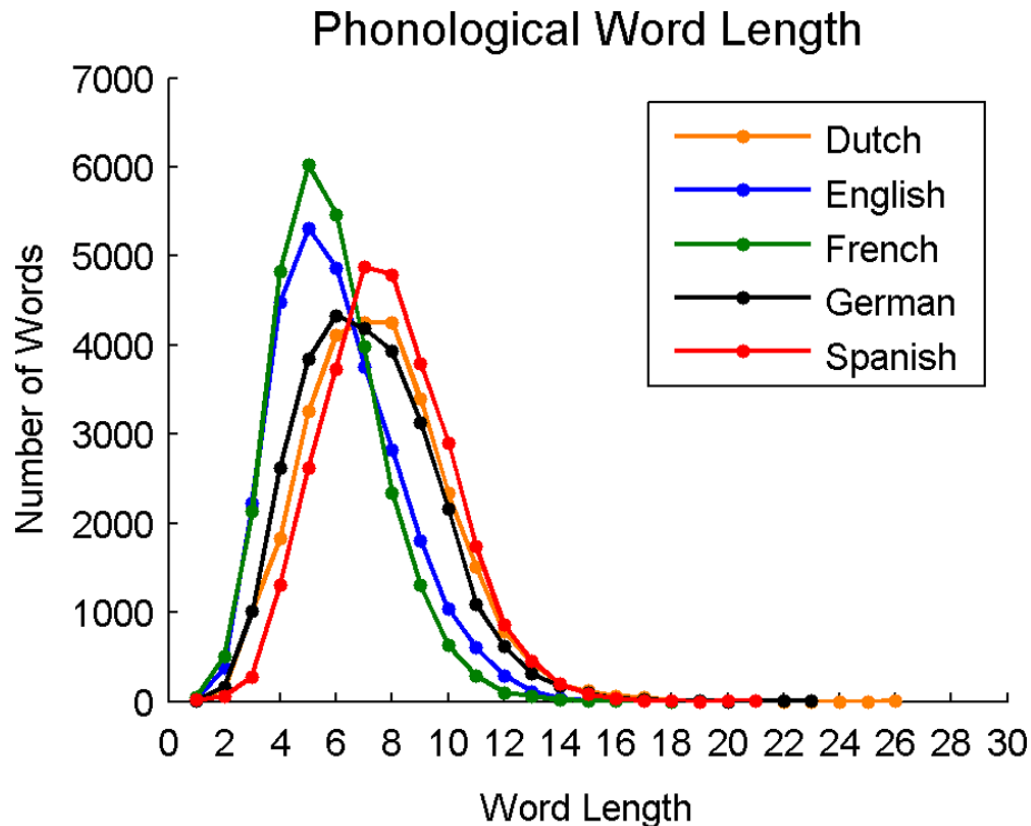
<b>Model</b>	<b>Sensitive to absolute position?</b>	<b>Probability type</b>	<b>Word boundary symbols?</b>	<b>Aggregation</b>	<b>Frequency Weighting</b>
<b><i>n</i>-gram</b>	No	Conditional	Yes	Product	Type
<b>V&amp;L</b>	Yes	Joint	No	Sum	Token



# Mathematical considerations

The V&L metric **does not**  
**define valid probabilities**

Tying probabilities to absolute  
position leads to **data**  
**scarcity** issues



# Linguistic considerations: translation invariance

Phonotactics is **translation invariant** (McCarthy & Prince 1986, Alderete et al. 2012, Hayes 2012)

- Linguistic models employ ‘constraints’ against certain structures
- Doesn’t matter where in the word a structure is located
- Ideas like “the 7<sup>th</sup> sound in a word” don’t seem to be necessary

V&L is not translation invariant; standard n-grams are

# Linguistic considerations: word-final restrictions

**Phonotactic restrictions related to word-final position are very common!**

- Dutch words cannot end in [d], [g], [b], ...
- Hawaiian words cannot end in a consonant

V&L can't refer to word-final position

- The index that corresponds to word-final position depends on length of word

The standard n-gram model can

# Linguistic considerations: word length

Phonological models often include a **preference for shorter words**

(e.g. Prince & Smolensky 1993, Goldwater et al. 2009, Pearl et al. 2010, Daland 2015, Johnson et al. 2018)

The use of addition in V&L imposes a preference for longer words

- Each additional sound raises the score

Standard n-grams prefer shorter words because they multiply probabilities

# Linguistic considerations: type vs. token weighting

Phonotactic generalizations are usually modeled based on **type frequency**

(e.g. Chomsky and Halle 1965, 1968, Pierrehumbert 2001, Bailey & Hahn 2001, Edwards et al. 2004, Mayer 2020)

V&L uses token frequencies

We train the standard n-gram model on type frequencies

# Model Bake-Off: Round 1 (Mayer, Kondur and Sundara, accepted)

Let's compare the standard n-gram and V&L models against eight publicly available phonotactic acceptability judgment datasets

**Question:** Which model predicts human responses the best?

# Datasets used in model comparison

Paper	Lang	Subjects	Stimuli	Input	Presentation
<b>Albright &amp; Hayes (2003)</b>	English	20	58 3-5 segment, monosyllabic nonce verbs	Likert scale	Auditory
<b>Daland et al. (2011)</b>	English	48	96 disyllabic nonce words differing in the initial onset	Likert scale	Orthographic
<b>Needle et al. (2022)</b>	English	1440	8400 nonce words, between 4-7 segments	Likert scale	Orthographic
<b>Scholes (1966)</b>	English	33	62 monosyllabic nonce words differing in initial onset	Forced choice	Orthographic
<b>Hayes &amp; White (2013)</b>	English	29	160 nonce words, between 2 and 7 segments	Magnitude estimation	Orthographic and auditory

# Datasets

<b>Paper</b>	<b>Lang</b>	<b>Subjects</b>	<b>Stimuli</b>	<b>Input</b>	<b>Presentation</b>
<b>Jarosz &amp; Rysling (2017)</b>	Polish	81	159 nonce words varying in onset properties	Likert scale	Orthographic
<b>Mayer &amp; Sundara (in prep)</b>	Spanish	168	575 CVCV nonce words	Magnitude estimation	Orthographic and auditory
<b>Mayer (in press)</b>	Turkish	90	596 CVCVC nonce words	Magnitude estimation	Orthographic and auditory



# Procedure for each dataset

1. Train each of the models on a representative training dataset
2. Use the trained models to score the stimuli from the study
3. Predict participant responses with a (linear/logistic) regression model

`response ~ uni_prob * bi_prob`

4. Compare models using the Akaike Information Criterion (AIC, Akaike 1974)

# AIC Rules of Thumb

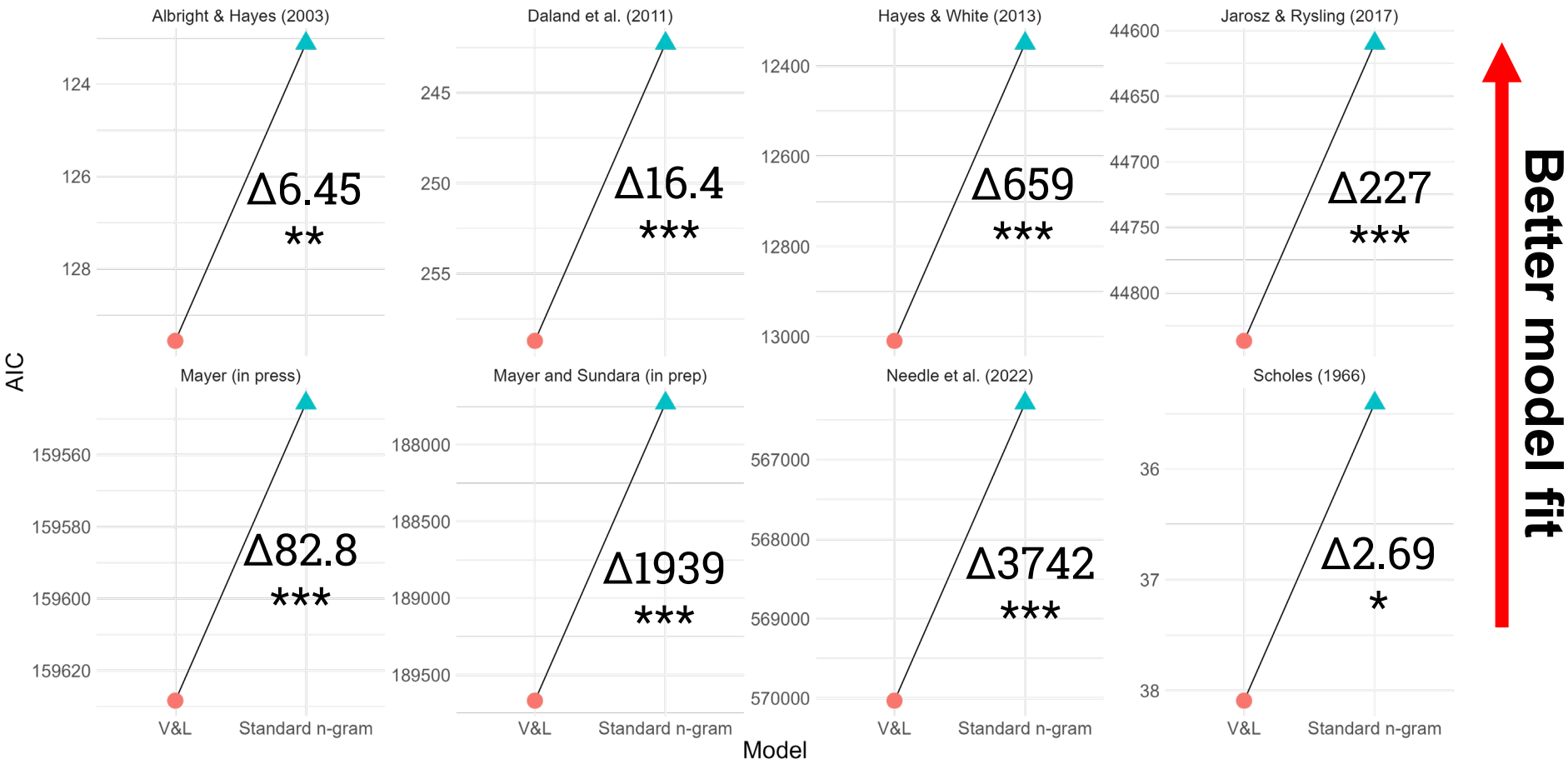
AIC is an estimate of prediction accuracy on held-out data

- We interpret AIC in terms of differences between models
- Lower AIC indicates better fit to data

We'll use a rule of thumb from Burnham and Anderson (2004)

- $\Delta AIC \leq 2$ : no difference between models
- $\Delta AIC > 10$ : strong support for model with lower AIC
- Increasing  $\Delta AIC$  indicates increasing certainty in better model

# Standard n-grams are better in every case



# Model Bake-Off 2: but *why*? (Mayer, Wagner & Sundara in prep)

The two models differ on five dimensions

Model	Sensitive to absolute position?	Probability type	Final word boundary?	Aggregation	Frequency Weighting
<i>n</i> -gram	No	Conditional	Yes	Product	Type
V&L	Yes	Joint	No	Sum	Token

Which of these are most important for the performance of the model?

# Token-weighting

Mayer, Kondur and Sundara (accepted) compared type- and token-weighted versions of both models

**Token weighting was almost always equal to or worse than type weighting**

- One dataset was fit better by token frequencies (Hayes and White 2013)

We'll only consider type-weighted models here

# Model Bake-Off 2: but *why*? (Mayer, Wagner & Sundara in prep)

We'll consider these four dimensions

Model	Sensitive to absolute position?	Probability type	Word boundaries	Aggregation
n-gram	No	Conditional	Yes	Product
V&L	Yes	Joint	No	Sum

Which of these are most important for the performance of the model?

# Bake-Off 2 procedure

We implemented 16 different models

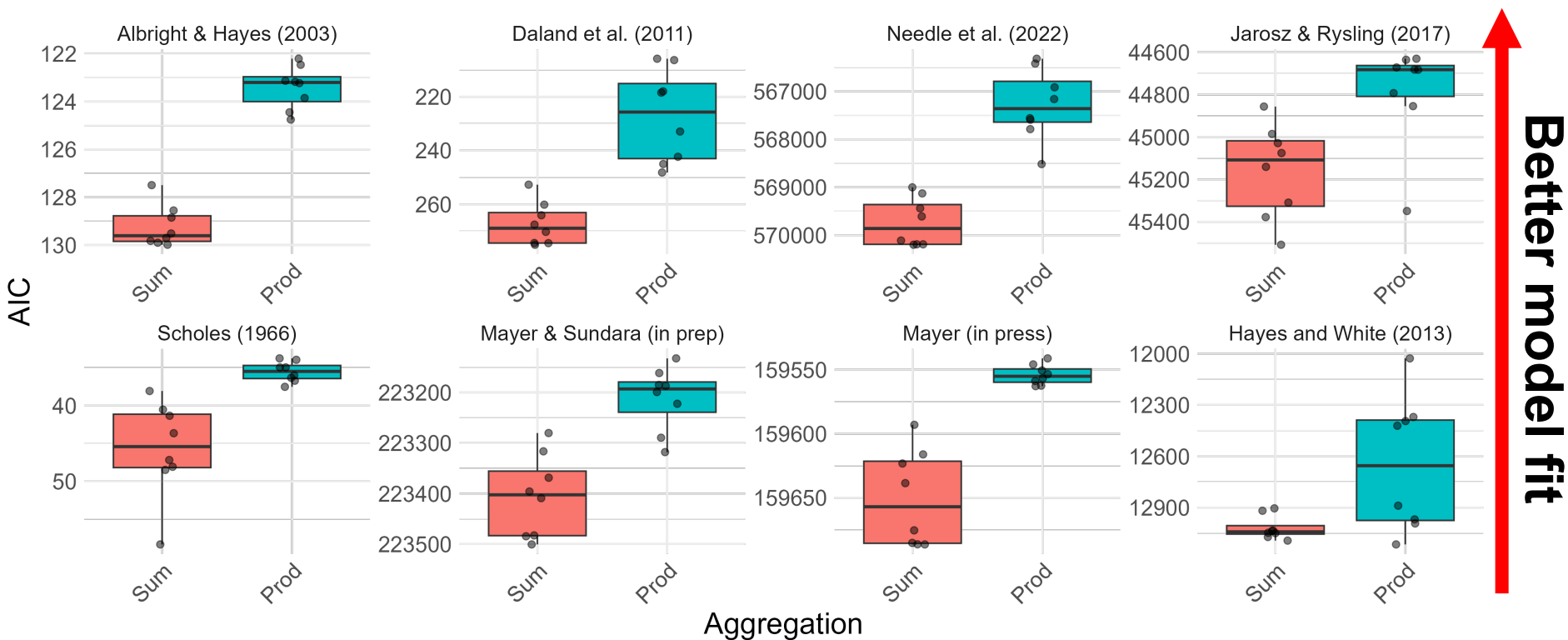
- One model per possible combination of the four parameters

$\{\text{joint, conditional}\} \times \{\text{relative, absolute}\} \times \{\text{sum, product}\} \times \{\text{no \#, yes \#}\}$

We fit each model to each dataset

- We'll compare models based on their values for each parameter

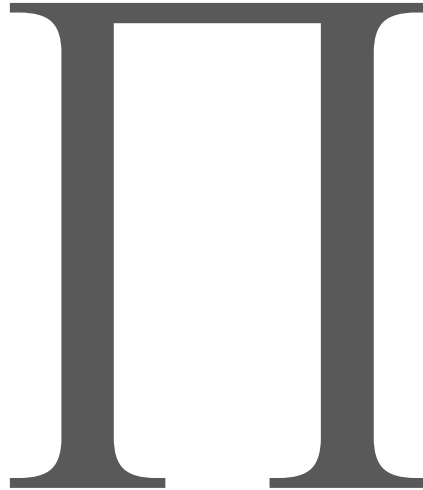
# Result 1: Adding probabilities is worse



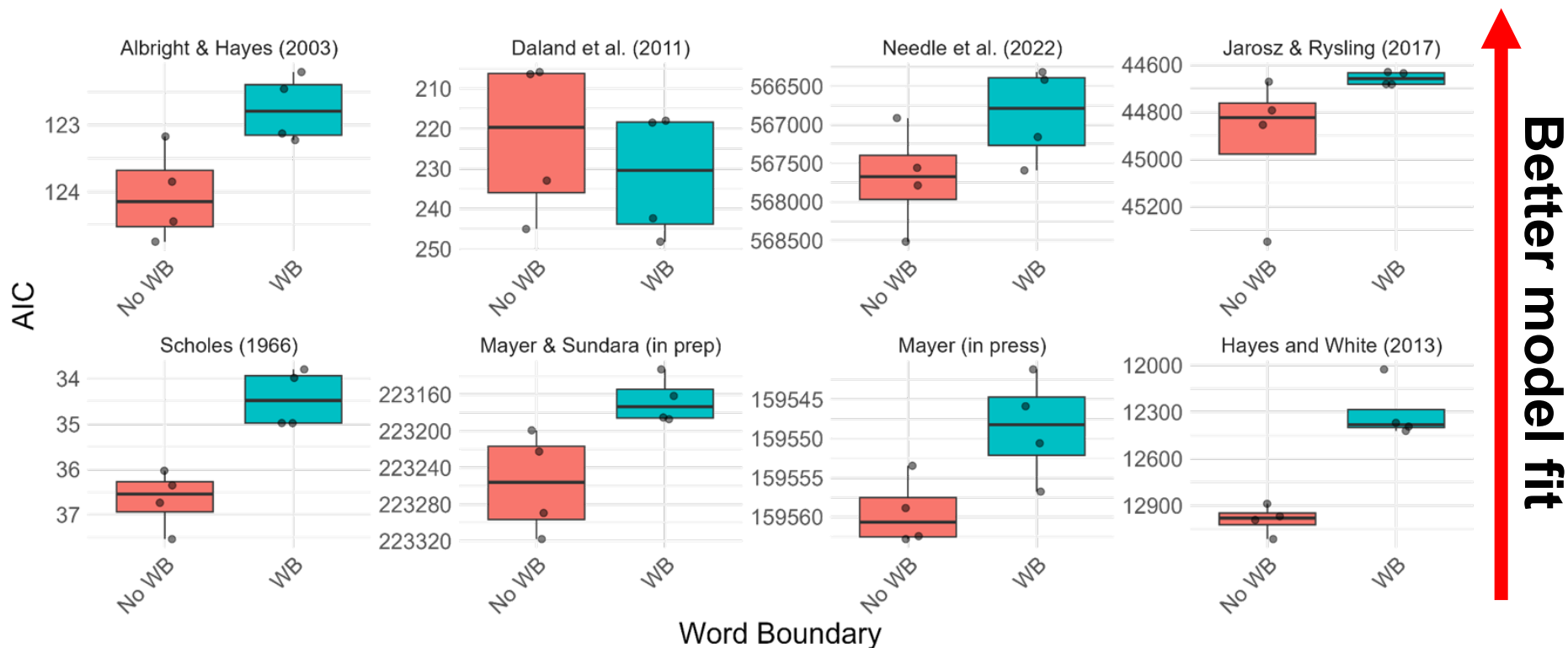
Each data point is an individual model



We'll only consider the 'product' models going forward

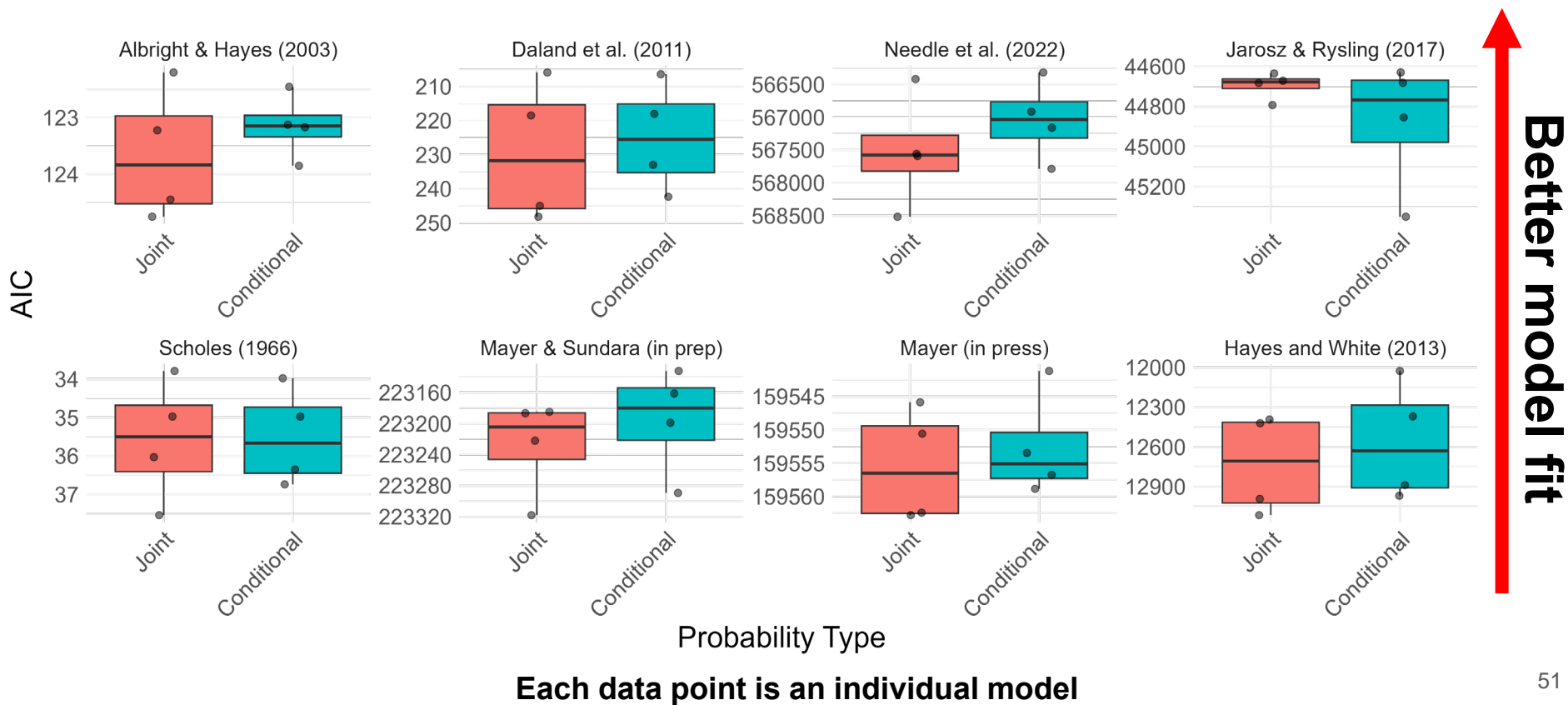


## Result 2: Strong preference for word boundaries

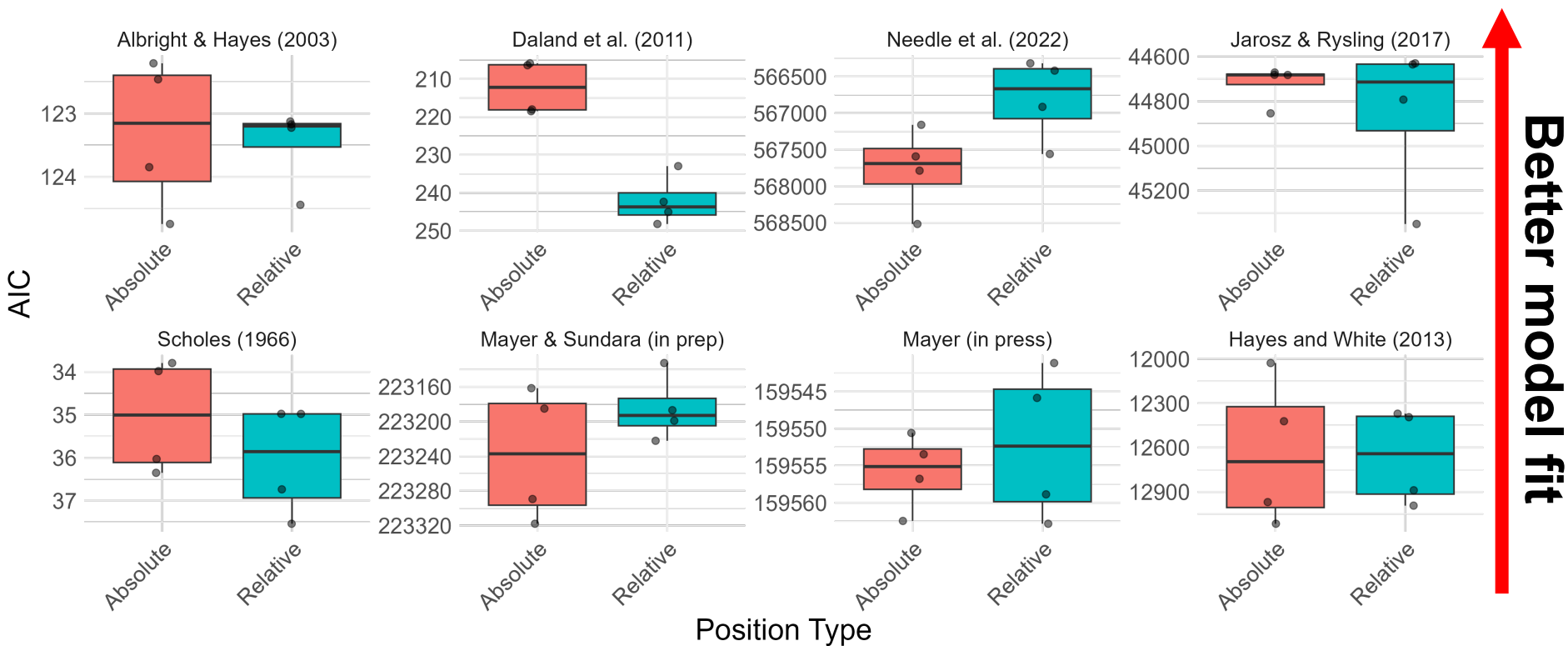


Each data point is an individual model

# Result 3: General preference for conditional probabilities



# Result 4: General preference for relative position



Each data point is an individual model

$$10 \geq \Delta AIC > 2$$

$$\Delta AIC > 10$$

## Bake-Off 2 Results

<u>Paper</u>	<u>Aggregation</u>	<u>Word Boundaries</u>	<u>Probability Type</u>	<u>Position Type</u>
Albright & Hayes (2003)	Prod > Sum	—	—	—
Daland et al. (2011)	<b>Prod &gt; Sum</b>	<b>No WB &gt; WB</b>	—	<b>Absolute &gt; Relative</b>
Jarosz & Rysling (2017)	<b>Prod &gt; Sum</b>	<b>WB &gt; No WB</b>	Conditional > Joint	<b>Relative &gt; Absolute</b>
Mayer (in press)	<b>Prod &gt; Sum</b>	<b>WB &gt; No WB</b>	Conditional > Joint	<b>Relative &gt; Absolute</b>
Mayer & Sundara (in prep)	<b>Prod &gt; Sum</b>	<b>WB &gt; No WB</b>	<b>Conditional &gt; Joint</b>	<b>Relative &gt; Absolute</b>
Needle et al. (2022)	<b>Prod &gt; Sum</b>	<b>WB &gt; No WB</b>	<b>Conditional &gt; Joint</b>	<b>Relative &gt; Absolute</b>
Scholes (1966)	Prod > Sum	WB > No WB	—	—
Hayes & White (2013)	<b>Prod &gt; Sum</b>	<b>WB &gt; No WB</b>	<b>Conditional &gt; Joint</b>	<b>Absolute &gt; Relative</b>

# What makes a good phonotactic model?

## Immediate practical consequence

V&L is worse than standard n-gram models across all the data sets we examined

- All five dimensions we compared favor standard n-grams

## Theoretical perspective

Standard n-grams align more closely with linguistic theory than V&L

- These results are a validation of these theoretical perspectives

# Relating model properties and linguistic theory

## 1. Combining probabilities with addition is a bad idea

- Probably reflects a bias towards shorter words

(e.g. Goldwater et al. 2009, Pearl et al. 2010, Daland 2015, Johnson et al. 2018)

## 2. Encoding word-final boundaries is important

- Humans are sensitive to structure at word edges!

(e.g. Monaghan and Christiansen 2010, Johnson et al. 2015, Sundara, Breiss, Dickson and Mayer under review)

# Zooming in on model properties

## 3. Conditional probabilities > joint probabilities

- The two are highly correlated (Gaygen 1997, Vitevitch and Luce 1999)
- Only conditional probabilities get us a valid probability distribution

## 4. Relative > absolute

- Phonotactics is translation invariant (McCarthy & Prince 1986, Alderete et al. 2012, Hayes 2012)
- Differences between datasets likely related to specific stimuli used
  - i. bigrams can't 'see' certain longer dependencies
  - ii. Positional model can track (some of) this information



# Roadmap

1. Background on phonotactics
2. Study 1: Theory comparison using phonotactic models
3. Study 2: Infant acquisition of phonotactics
4. Discussion and take-aways

## Study 2: Phonotactics and word learning

Sundara, Breiss, Dickson & Mayer (submitted). *Developmental Science*.



Megha Sundara (UCLA)



Canaan Breiss (USC)



Niels Dickson (UCI)

# A puzzle

We've long known infants are sensitive to phonotactics at 8 months

(Jusczyk et al., 1994; Thiessen & Erickson, 2013; Sundara et al., 2022)

- Also at 5 months (Sundara & Breiss resubmitted)

**Problem:** 5-month-olds don't "know" many words (~20; Bergelson & Swingley 2011)

Where does infants' phonotactic knowledge come from?

- What's in the lexicon?

# Hypothesis 1: The prelexical hypothesis

Infants learn phonotactic generalization from unsegmented speech

(e.g., Adriaans & Kager, 2010; Brent & Cartwright, 1996; Daland & Pierrehumbert, 2011)

- Infants access word boundary information from utterance edges

(Christophe et al. 1997; Johnson & Seidl, 2008)

- Computationally simple



# Hypothesis 2: The strong lexical hypothesis

Infants learn phonotactic generalization from words associated with referents

(Sundara and Breiss resubmitted)

- Aligns with our characterization of adult phonotactic generalization



# Hypothesis 3: The protolexical hypothesis

Infants learn phonotactics from word forms that need not be associated with referents (Jusczyk, Houston & Newsome, 1999; Ngon et al., 2011; Kim & Sundara 2021)

- Infants can segment speech by 5 months

(Thiessen & Erickson, 2013;  
Johnson & Tyler, 2010)



# How do we test this?



Sundara & Breiss (resubmitted) tested 5-mo-olds' ability to discriminate between word forms with different phonotactic probabilities

Stimuli were chosen based on adult norming data

- Total of 396 fake words with Consonant-Vowel-Consonant structure
- Varied in their **unigram and bigram probabilities** (from V&L model)

# Infant experiments (Sundara & Breiss, resubmitted)

Monolingual English learning 5-month-olds

- > 90% exposure to English

Three experiments

- 2a: High vs. low unigram probability, low bigram probability (**n=30**)
- 2b: (Less) high vs. low unigram probability, low bigram probability (**n=30**)
- 2c: High vs. low unigram and bigram probability (**n=38**)

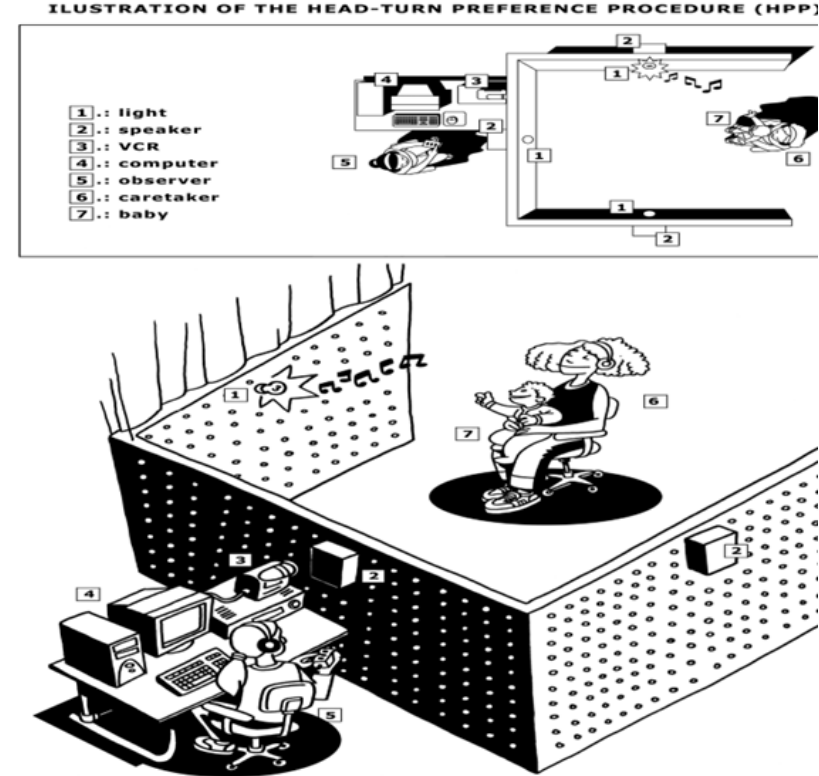


# Method

Experiments used Headturn Preference Procedure, following Juscyk et al. (1994)

Completely infant-controlled preference experiment

- 2 familiarization trials with music
- 12 test trials, low vs. high probability items



# Results

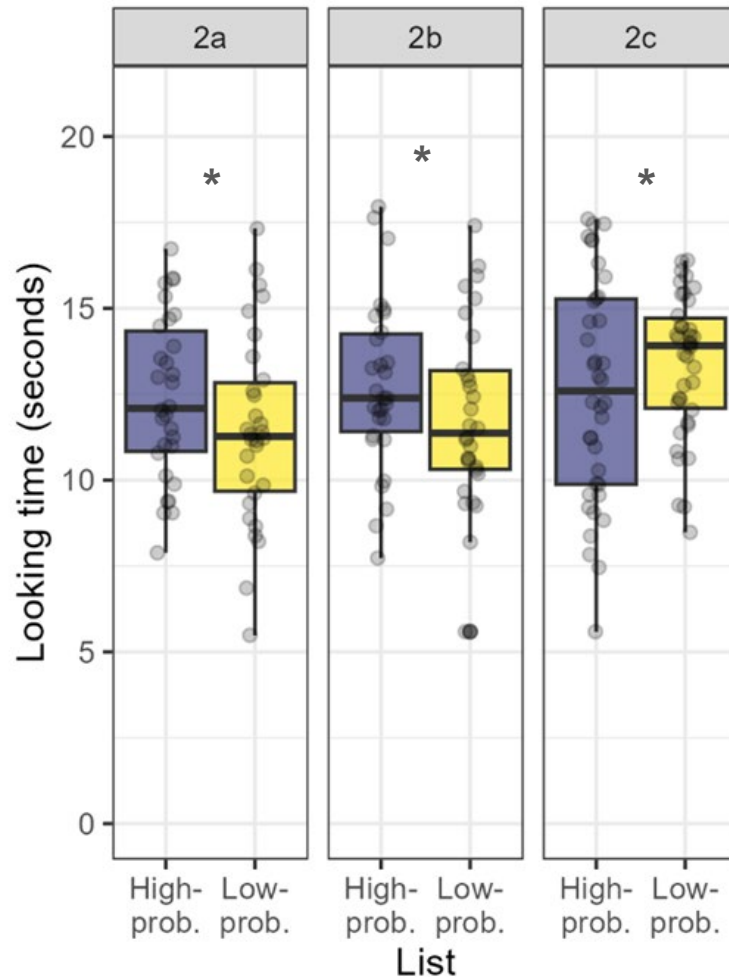
English learning 5-mo-olds are sensitive to segmental dependencies

Have both cues in 2c makes it easier for infants!

- And results in novelty preference

(Hunter & Ames 1988)

We now have three stimulus sets that 5-mo-olds can distinguish



# Testing hypotheses about phonotactic learning

Approach:

1. Create a **corpus** embodying each hypothesis
2. Fit **unigram and bigram models** to corpus (V&L)
3. Use model to **score experimental stimuli** for unigram/bigram probability
4. Test if assigned probabilities **distinguish** high vs. low probability stimuli

**Novel aspect:** we're comparing model performance against infant behavior rather than adult performance.

# 1: Prelexical hypothesis

Infants learn phonotactics from unparsed utterances

(e.g., Adriaans & Kager, 2010; Brent & Cartwright, 1996; Daland & Pierrehumbert, 2011)

**Corpus**: 15,527 utterances (types) with no word boundaries from Pearl-Brent corpus of infant-directed speech (phonetically transcribed)

#noeatingdogfood#  
#theresmorgansbook#  
#ohnoonewantstogettdressed#

## 2: Strong lexical hypothesis

Infants learn phonotactics from words with associated referents

At 5-months, infants associate some word forms with referents

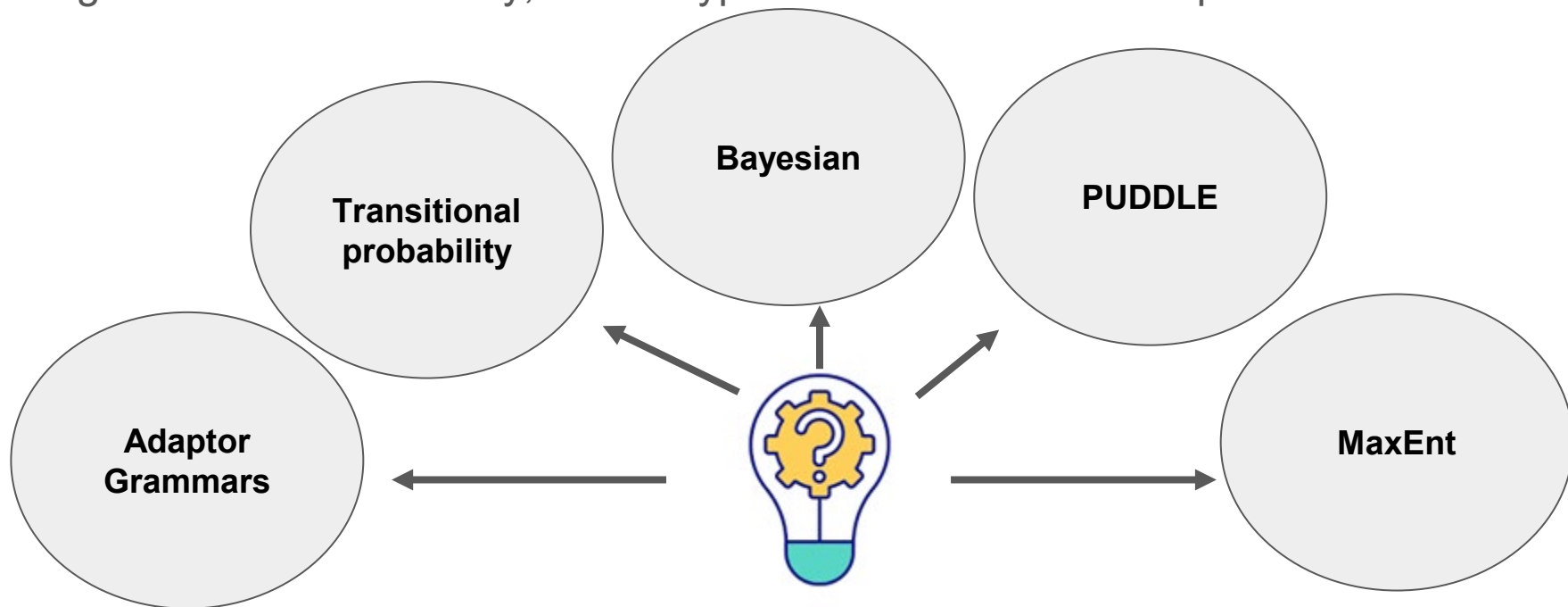
(Bergelson & Swingley, 2012; Bortfeld et al., 2005)

- *ear, eyes, face, foot, feet, hair, hand(s), leg(s), mouth, nose, apple, banana, bottle, cookie, juice, milk, spoon, yogurt* (Bergelson & Swingley 2011), *mommy, daddy* (Bortfeld et al. 2005)

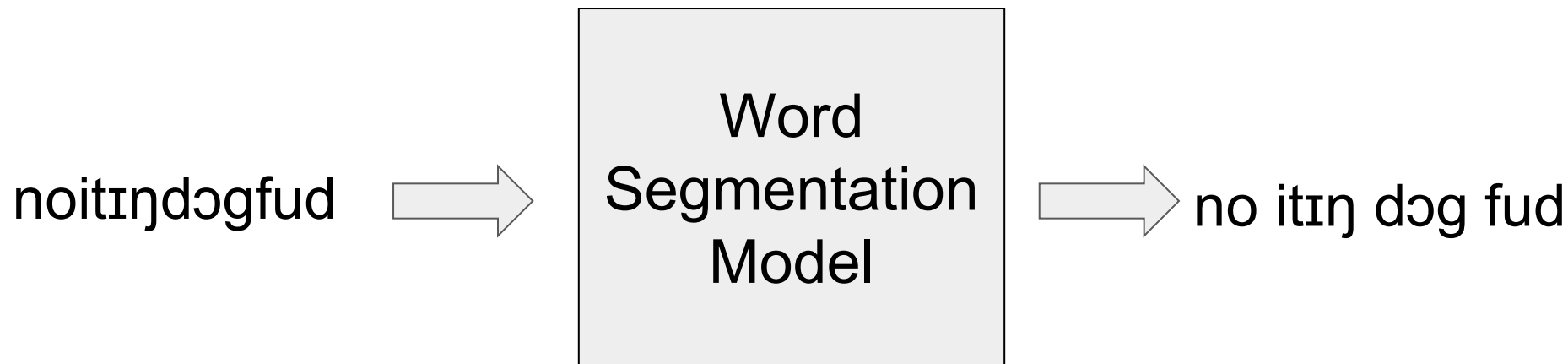
**Corpus**: 18 stems; 22 words

### 3: Protolexical hypothesis

**The premise:** The output of any *unsupervised model of word segmentation*, regardless of its accuracy, is one hypothesis about the infant proto-lexicon



# Word segmentation



# Categorizing word segmentation models

**Lexicon-based:** does the model use previously identified words to segment future utterances?





# Categorizing word segmentation models

**Phonotactics-based:** does the model evaluate the likeliness of a segmented word based on its phonotactic properties?



✗ Lexicon-based  
✗ Phonotactics-based

*Baseline models*

✓ Lexicon-based  
✗ Phonotactics-based

*Bayesian Models*

✗ Lexicon-based  
✓ Phonotactics-based

*Transitional Probability  
Models*

✓ Lexicon-based  
✓ Phonotactics-based

*Adaptor Grammars*

*PUDDLE*

*MaxEnt*

### 3: Protolexical hypothesis

Infants learn phonotactics from word forms in the lexicon

(Thiessen, Kronstein & Huffnagle, 2013)

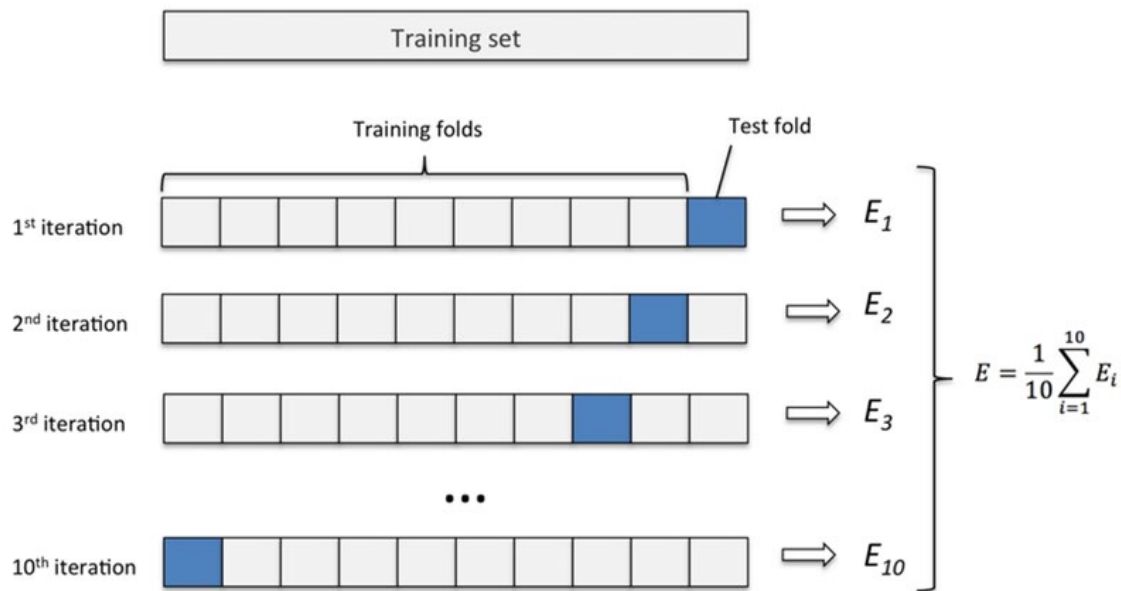
**Corpus**: Output of 24 unsupervised models of word segmentation on Pearl-Brent corpus of infant-directed speech

- 24 distinct hypotheses about word segmentation strategies

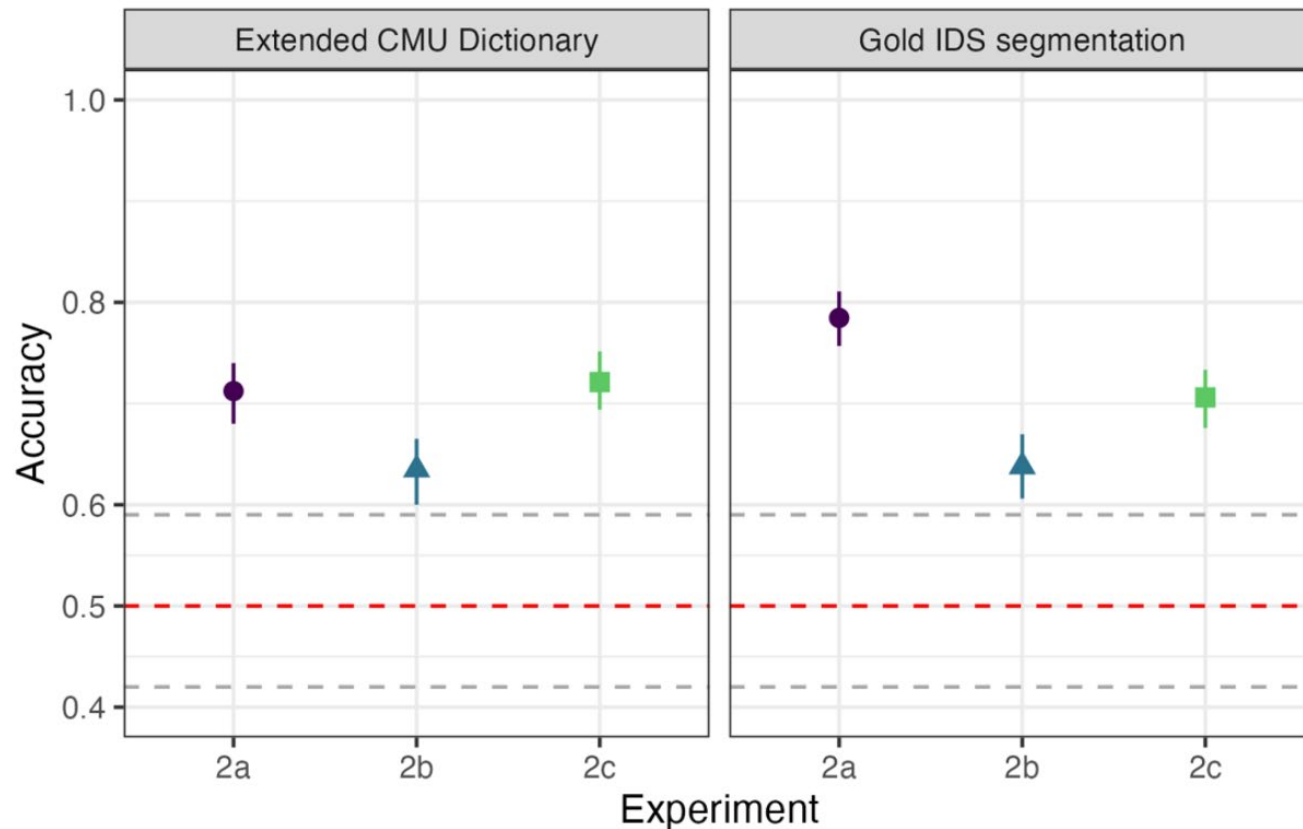
Mostly run using wordseg (Bernard et al. 2019)

# Relating word scores and infant behavior

High vs. low probability word  $\sim$  unigram\_probability \* bigram\_probability



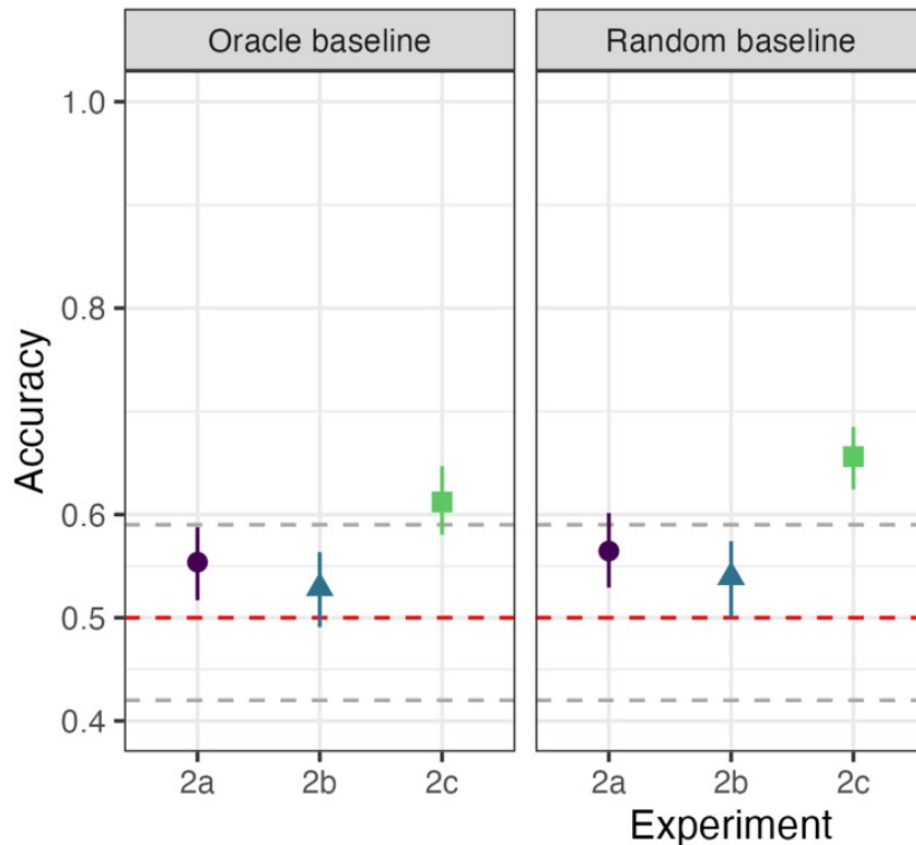
# Sanity Check



Adult lexicons & fully-segmented infant-directed speech provide sufficient information to distinguish lists distinguished by 5-month-olds.

✗ Lexicon-based  
✗ Phonotactics-based

Both baselines provide sufficient  
information to distinguish list 2c!



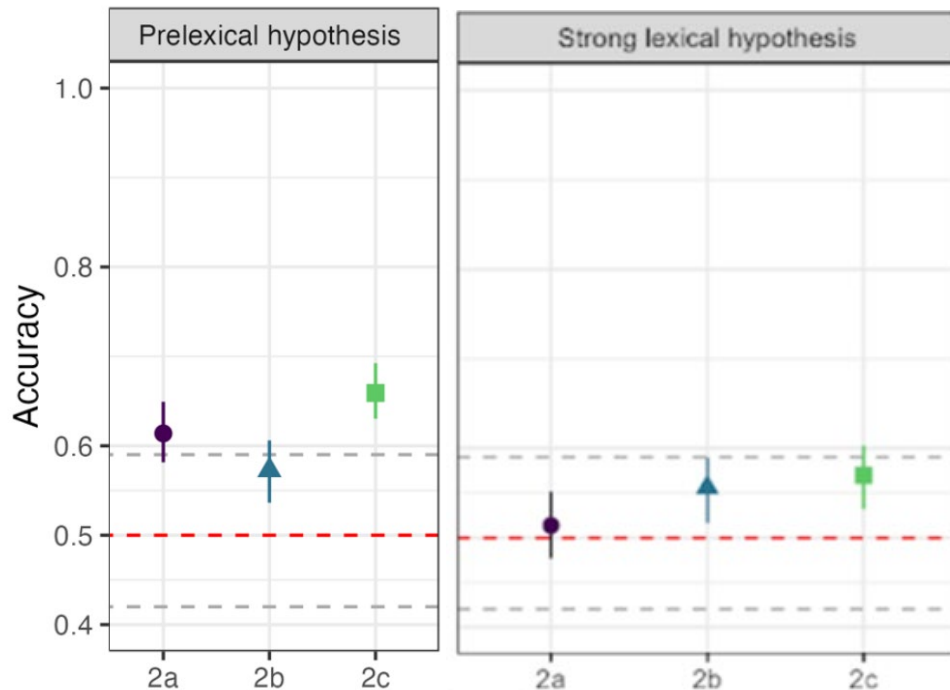
# Prelexical and Strong Lexical Hypotheses

## Prelexical hypothesis

Distinguishes 2a and 2c, but not 2b

## Strong lexical hypothesis

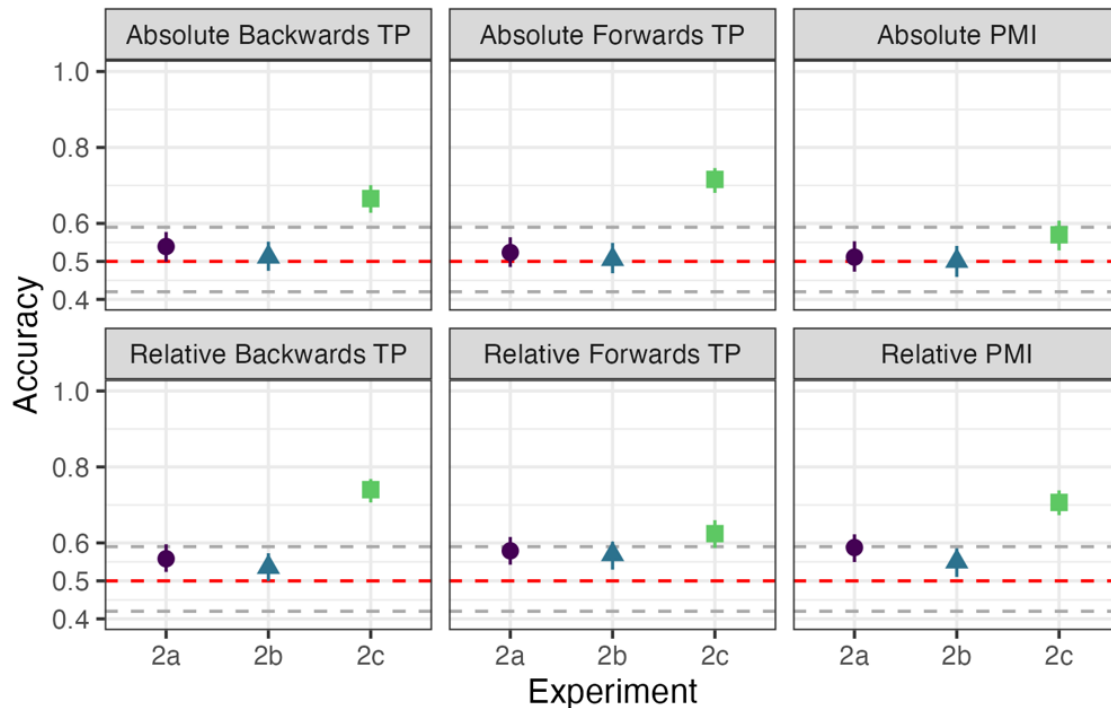
Worse than guessing randomly



✗ Lexicon-based  
✓ Phonotactics-based

0/6 distinguish all three lists

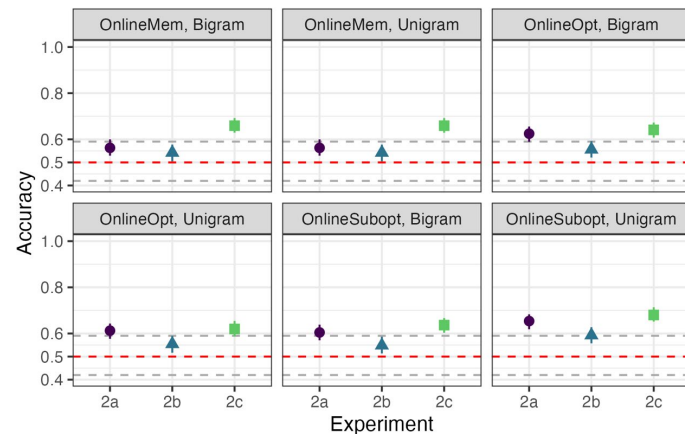
Best model = Baseline



Transitional Probability-based models (Saksida et al. 2017)

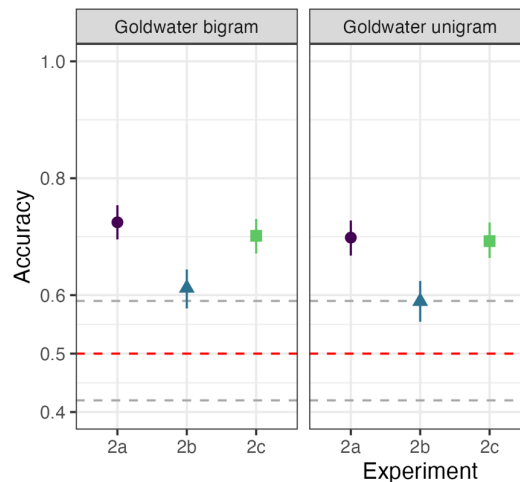


✓ Lexicon-based  
✗ Phonotactics-based



Phillips & Pearl (2015)

1/8 models distinguish all lists

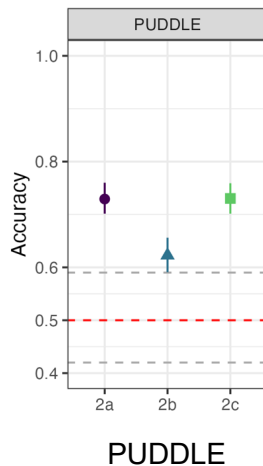


Goldwater et al. (2009)

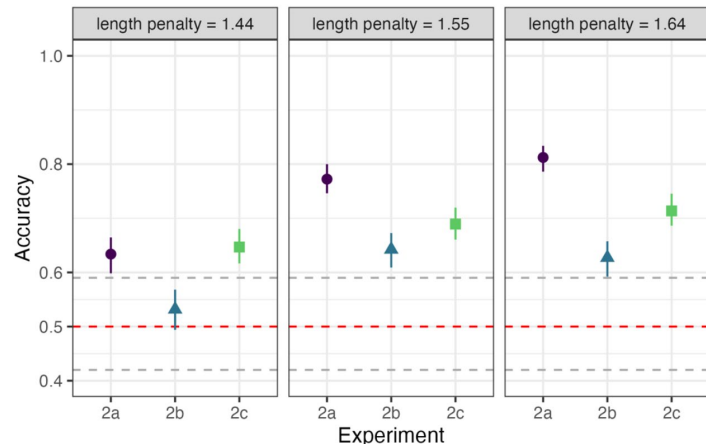
- ✓ Lexicon-based
- ✓ Phonotactics-based

## 7/10 models distinguish all three lists

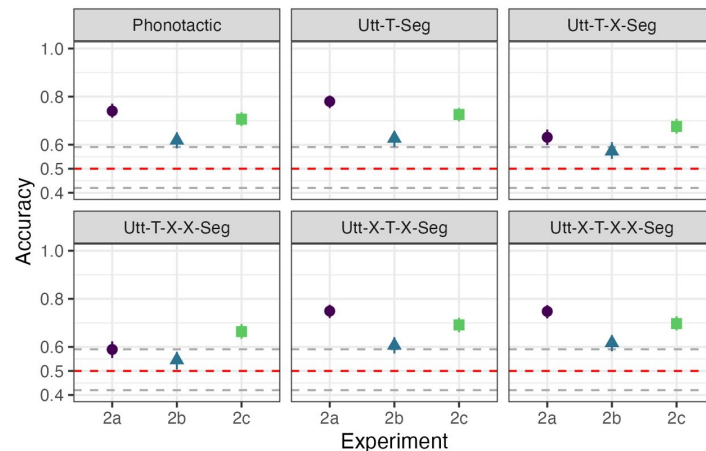
- MaxEnt: 2/3
- Adaptor grammar: 4/6
- PUDDLE: 1/1



(Monaghan and Christiansen 2010)




MaxEnt models (Johnson, Pater, Staubs & Dupoux, 2015)



Adaptor grammar models (Johnson et al. 2006)

# Evaluating mechanisms

5-month-olds' sensitivity to phonotactic patterns is predicted by

- Prelexical hypothesis 
- Strong lexical hypothesis 
- Protolexical hypothesis  (some proposals)

Successful protolexical models rely on stored words to bootstrap future segmentation and apply phonotactic restrictions to segmentation.

Caveat: All protolexical hypotheses are better at segmenting words than 5-month-olds!

# Are successful models the best segmenters?

Model and source	Word segmentation F-score
JPSD Maxent (Johnson et al. 2015), $d = 1.55$	0.86
Adaptor Grammar, Phonotactic	0.78
JPSD Maxent (Johnson et al. 2015), $d = 1.64$	0.76
Adaptor Grammar, U-T-Seg (see main text)	0.75
PUDDLE (Monaghan et al. 2012)	0.72
JPSD Maxent (Johnson et al. 2015), $d = 1.44$	0.67
Adaptor Grammar, U-X-T-X-X-Seg	0.66
BatchOpt, unigram (Goldwater et al. 2009)	0.63
BatchOpt, bigram (Goldwater et al. 2009)	0.63
Adaptor Grammar, U-X-T-X-Seg	0.62
Adaptor Grammar, U-T-X-Seg	0.61
Adaptor Grammar, U-T-X-X-Seg	0.45
Oracle baseline	0.26
Random baseline	0.10

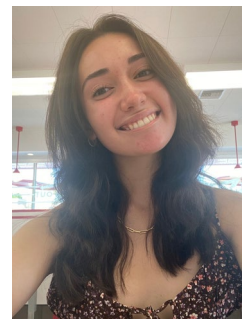
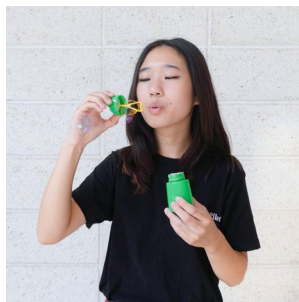
Not always!

# Future directions

The role of prosody:

- Infants are sensitive to large prosodic boundaries
- Is prosodic information *within* the utterance sufficient for phonotactic learning at 5-mo?

Work in progress with Will Chang and undergraduate RAs Alison Howland and Lauren Hsu



# Future directions

## Comparison across languages

- Are the same segmentation strategies applicable in languages with different morphophonology?
- We've collected norming data on Spanish adults (Mayer et al. 2024)
- Spanish infant study to come

# Roadmap

1. Background on phonotactics
2. Study 1: Theory comparison using phonotactic models
3. Study 2: Infant acquisition of phonotactics
4. Discussion and take-aways

# What have we learned?

These two studies focused on separate aspects of phonotactic learning

- But both take the same broad approach

Model comparison helps us understand how how phonotactic learning progresses

**There's no such thing as a theory-neutral model!**



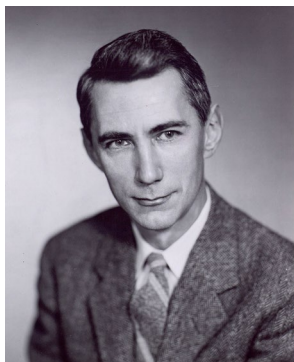
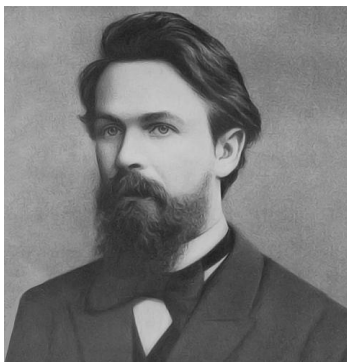
# Study 1: Comparing phonotactic models

The standard n-gram model most consistently predicts experimental responses

- Validates several claims of linguistic theory

**Caveat:** n-grams are an insufficient (but useful!) model of phonotactics

- More complex models will probably preserve these useful properties



## Study 2: Infant learning of phonotactics

Modeling work supports the **protolexical hypothesis**: infants learn phonotactic generalizations from hypothesized word forms

Word segmentation models best support infant phonotactic generalizations when:

1. They use previously identified words to bootstrap segmentation
2. They evaluate possible new words based on identified phonotactic restrictions

# Sharing is caring

We were able to undertake both of these studies because researchers made their code and datasets publicly available

- This is a big part of the popularity of the V&L model

Our code and data are available for reference and reuse (see papers)

# The UCI Phonotactic Calculator (Mayer, Kondur and Sundara, accepted)

[Home](#) [About](#) [Datasets](#) [GitHub](#)

## UCI Phonotactic Calculator

### Welcome to the UCI Phonotactic Calculator!

This is a research tool that allows users to calculate a variety of *phonotactic metrics*. These metrics are intended to capture how probable a word is based on the sounds it contains and the order in which those sounds are sequenced. For example, a nonce word like [stik] 'steek' might have a relatively high phonotactic score in English even though it is not a real word, because there are many words that begin with [st], end with [ik], and so on. In Spanish, however, this word would have a low score because there are no Spanish words that begin with the sequence [st]. A sensitivity to the phonotactic constraints of one's language(s) is an important component of linguistic competence, and the various metrics computed by this tool instantiate different models of how this sensitivity is operationalized.

The general use case for this tool is as follows:

1. Choose a *training file*. You can either upload your own or choose one of the default training files (see the [About](#) page for details on how these should be formatted and the [Datasets](#) page for a description of the default files). This file is intended to represent the input over which phonotactic generalizations are formed, and will typically be something like a dictionary (a large list of word types). The models used to calculate the phonotactic metrics will be fit to this data.
2. Upload a *test file*. The trained models will assign scores for each metric to the words in this file. This file may duplicate data in the training file (if you are interested in the scores assigned to existing words) or not (if you are interested in the predictions the various models make about how speakers generalize to new forms).

The calculator computes a suite of metrics that are based on unigram/bigram frequencies (that is, the frequencies of individual sounds and the frequencies of adjacent pairs of sounds). This includes type- and token-weighted variants of the positional unigram/bigram method from Jusczyk et al. (1994) and Vitevitch and Luce (2004), as well as type- and token-weighted variants of standard unigram/bigram probabilities. See the [About](#) page for a detailed description of how these models differ and how to interpret the scores.

The UCI Phonotactic Calculator was developed by [Connor Mayer](#) (UCI), Arya Kondur (UCI), and [Megha Sundara](#) (UCLA). Please direct all inquiries to Connor Mayer ([cjmayer@uci.edu](mailto:cjmayer@uci.edu)).

### Citing the UCI Phonotactic Calculator

If you publish work that uses the UCI Phonotactic Calculator, please cite the GitHub repository:

Mayer, C., Kondur, A., & Sundara, M. (2022). UCI Phonotactic Calculator (Version 0.1.0) [Computer software]. <https://doi.org/10.5281/zenodo.7443706>

## Provide Input for Calculations

### Upload a training file or select a default file

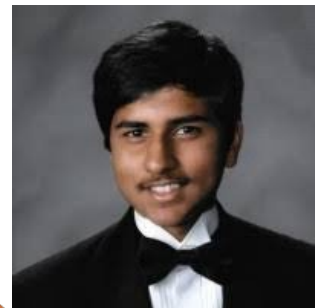
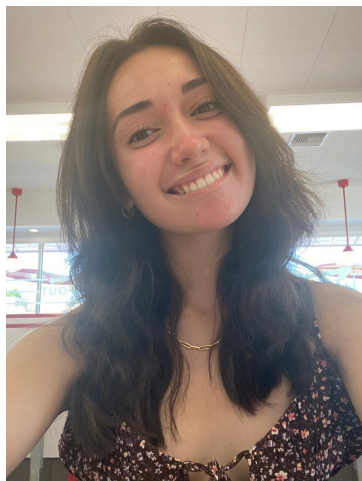
Training file:  No file selected.

Default training file:

Test file:  No file selected.

<https://phonotactics.socsci.uci.edu/>

# Thank you!



# References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723.
- Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1), 9-41.
- Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2), 119-161.
- Bailey, T.M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language* 44:568–591.
- Burnham, K.P., & Anderson, D.R. (2004). Multimodal inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2): 261-304.
- Bybee, J. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, 10(5), 425-455.
- Bybee, J. (2003). *Phonology and language use* (Vol. 94). Cambridge University Press.
- Castro, N. & Vitevitch, M.S. (2023). Using Network Science and Psycholinguistic Megastudies to Examine the Dimensions of Phonological Similarity. *Language and speech*, 66(1), 143–174.

# References

- Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, 1(2):97–138.
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
- Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In Coleman, J. (ed.), *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology*. Association for Computational Linguistics, Somerset, NJ: 49-56.
- Dai, H., Mayer, C., & Futrell, R. (2023). Rethinking representations: A log-bilinear model of phonotactics. *Proceedings of the Society for Computation in Linguistics*, 6.
- Daland, R. (2015). Long words in maximum entropy phonotactic grammars. *Phonology*, 32(3), 353-383.
- Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Normann, I. (2011). Explaining sonority projection effects. *Phonology*, 28: 197–234.
- Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Interaction*.
- Gaygen, D. E. (1997). *The effects of probabilistic phonotactics on the segmentation of continuous speech*. Unpublished doctoral dissertation, SUNY, Buffalo.

# References

- Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition*, 107(3), 1155-1164.
- Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112, 21–54.
- Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, 44:45-75.
- Hayes, B., & Wilson, C. (2008) A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39, 379-440.
- Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in infancy research*.
- Jarosz, G., & Rysling, A. (2017). Sonority Sequencing in Polish: the Combined Roles of Prior Bias and Experience. *Proceedings of the 2016 Annual Meetings on Phonology, USC*.
- Johnson, M., Pater, J., Staubs, R., & Dupoux, E. (2015). Sign constraints on feature weights improve a joint model of word segmentation and phonology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 303–313).
- Marian, V., Bartolotti, J., Chabal, S., Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE* 7(8): e43230.



# References

Markov, A.A. (1913). Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain ('Example of a statistical investigation of the text of "Eugene Onegin" illustrating the dependence between samples in chain'). Izvestia Imperatorskoi Akademii Nauk (Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg), 7:153–162.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, 38(4), 465-494.

Mayer, C. (in press). Reconciling categorical and gradient models of phonotactics. *Proceedings of the Society for Computation in Linguistics*.

Mayer, C., Kondur, A., & Sundara, M. (resubmitted). The UCI Phonotactic Calculator: An online tool for computing phonotactic metrics. *Behavior Research Methods*.

Mayer, C., & Nelson, M. (2020). Phonotactic learning with neural language models. *Society for Computation in Linguistics*, 3(1).

Mayer, C., & Sundara, M. (in prep). Probing the phonotactic knowledge of Spanish-learning infants.

Mirea, N., & Bicknell, K. (2019, July). Using LSTMs to assess the obligatoriness of phonological distinctive features for phonotactic learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1595-1605).

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, 37(3), 545–564.

# References

- Needle, J. M., Pierrehumbert, J. B., & Hay, J. B. (2022). Phonotactic and Morphological Effects in the Acceptability of Pseudowords. In A. Sims, A. Ussishkin, J. Parker, & S. Wray (Eds.), *Morphological Diversity and Linguistic Cognition*. CUP.
- Norris, D. & McQueen, J.M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115: 357
- Pearl, L., Goldwater, S., & Steyvers, M. (2010). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, 8(2–3), 107–132.
- Pierrehumbert, J. (2001). Stochastic phonology. *Glott international*, 5(6), 195-207.
- Scholes, R. (1966). *Phonotactic grammaticality*. The Hague: Mouton.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379-423.
- Steffman, J., & Sundara, M. (2024). Disentangling the role of biphone probability from neighborhood density in the perception of nonwords. *Language & Speech*, 67 (1), 166-202.
- Sundara, M., Breiss, C., Dickson, N., & Mayer, C. (under review). What's in a 5-month-old's (proto-)lexicon? *Developmental Science*.

# References

Taylor, C. F., & Houghton, G. (2005). Learning artificial phonotactic constraints: time course, durability, and relationship to natural constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6), 1398.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3): 374–408.

Vitevitch, M.S., & Luce, P.A. (2004) A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers*, 36: 481-487.

# Appendices

# The UCI Phonotactic Calculator (Mayer, Kondur and Sundara, resubmitted)

The UCIPC is a website for computing a suite of phonotactic metrics

- Can be run using 10 built-in training sets across 7 languages
- Users can specify their own training data
- Trained models are used to score user-provided test data

The UCIPC computes

- Standard unigram and bigram probabilities
- PPC unigram and bigram probabilities
- Token-weighted and smoothed variants of each

# Training file

1	EY	633517.5
2	AHBAEK	59
3	AEBAHKAHS	8
4	AHBAENDAHN	1010
5	AHBAESH	15
6	AHBEYT	42
7	AEBIY	7
8	AEBEY	7
9	AEBIY	181
10	AEBAHT	43
11	AHBRIYVIEYET	35
12	AHBRIYVIEYSHAHN	14
13	AEBDAHKEYT	40
14	AEBDIHKEYSHAHN	34
15	AEBDOWMAHN	57
16	AEBDAHMAHN	57
17	AEBDAAMAHNAHL	63
18	AHBDAAMAHAHNAHL	63
19	AEBDAHKT	19
20	AEBDAHKSHAHN	5.5
21	AHBDAHKSHAHN	5.5
22	AHBEHD	4
23	AEBEHRAHNT	11
24	AEBEREYSHAHN	50
25	AHBEHT	33
26	AHBEYAHNS	17
27	AEBHHAOR	39
28	AHBHHAORAHNS	7
29	AEBHHAORAHNT	23
30	AHBAYD	84
31	AHBILAHATIY	1557
32	AEBJHEHKT	57
33	AHBLEYZ	29
34	EYBAHL	5887
35	AEBNAORMAHL	105
36	AEBNAORMAE LAHTIY	39
37	AA BOW	6
38	AHBAORD	285
39	AH BOWD	31
40	AHBAALIHSH	301



# Scored test file

1	word	word_len	uni_prob	uni_prob_freq_weighted	uni_prob_smoothed	uni_prob_freq_weighted_smoothed
2	BLIYG I H F	6	-21.28560225	-21.28547321	-21.36475687	-21.36471595
3	BLEH Z I H G	6	-21.89701032	-21.89653607	-21.96285725	-21.96272277
4	BRIYG I H F	6	-21.26431799	-21.26419239	-21.31293023	-21.31289144
5	BREH P I H D	6	-19.85093399	-19.85144946	-19.78328505	-19.78342243
6	BWIYG I H F	6	-23.46505863	-23.46365267	-23.44272982	-23.44239313
7	BWAAS I H P	6	-21.82616145	-21.82539077	-21.76996196	-21.76979186
8	DGEH P I H D	6	-20.91194316	-20.91206033	-20.85977901	-20.85980997
9	DGAAT I H F	6	-21.1446086	-21.14449317	-21.17316921	-21.17313346
10	DN IYG I H F	6	-20.55196506	-20.55203056	-20.5815925	-20.58160607
11	DNAAT I H F	6	-19.37124649	-19.37172047	-19.36533752	-19.36546055
12	DRIYG I H F	6	-20.8320401	-20.83206664	-20.83634114	-20.83634568
13	DREH P I H D	6	-19.4186561	-19.41932371	-19.30669597	-19.30687668
14	DWEH Z I H G	6	-23.64418881	-23.64258978	-23.56424112	-23.5638542
15	DWAAT I H F	6	-21.85206218	-21.85121683	-21.74988576	-21.74970185
16	FLEH Z I H G	6	-22.0996585	-22.09908688	-22.12310698	-22.12295264
17	FLAAT I H F	6	-20.30753186	-20.30771393	-20.30875163	-20.30880029
18	FN IYB I H D	6	-20.05862089	-20.05896282	-20.07368218	-20.07377139
19	FNEH Z I H G	6	-21.7982992	-21.79776998	-21.81653169	-21.81638852
20	FREH P I H D	6	-20.05358216	-20.05400026	-19.94353478	-19.9436523
21	FRAAS I H P	6	-19.82806898	-19.82848129	-19.80041209	-19.80052004
22	FWIYB I H D	6	-22.53943657	-22.53845917	-22.45823042	-22.4580127
23	FWEH Z I H G	6	-24.27911488	-24.27726633	-24.20107993	-24.20062982
24	GLEH P I H D	6	-20.36556242	-20.36579804	-20.34302202	-20.34308162
25	GLAAT I H F	6	-20.59822785	-20.59823087	-20.65641222	-20.6564051
26	GRIYB I H D	6	-20.62939192	-20.62951583	-20.67609141	-20.67611582
27	GRAAT I H F	6	-20.57694359	-20.57695004	-20.60458557	-20.60458059
28	GW IYB I H D	6	-22.83013257	-22.82897612	-22.80589101	-22.80561751
29	GWAA T I H F	6	-22.77768423	-22.77641033	-22.73438516	-22.73408229

# A plot from Hayes (2012)

