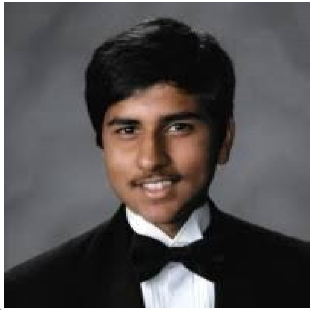
What's in a word? Using computational modeling to study phonotactic learning

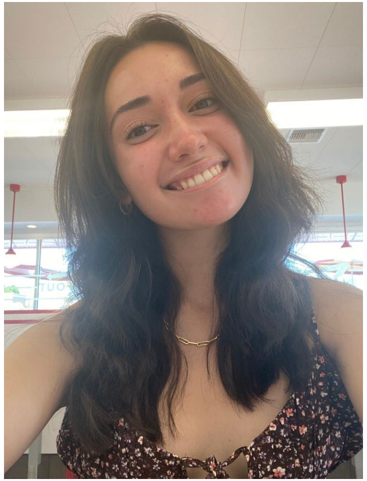
Connor Mayer
UCI Department of Language Science

CSUF Linguistics Symposium

Collaborators

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





Roadmap

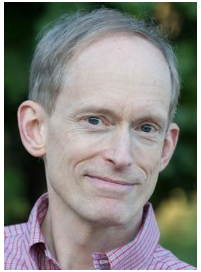
1. Why computational modeling?
2. Background on phonotactics
3. Relating phonotactic learning and word learning
4. Phonotactic model bake-off
5. Discussion and take-aways

Roadmap

1. Why computational modeling?
2. Background on phonotactics
3. Relating phonotactic learning and word learning
4. Phonotactic model bake-off
5. Discussion and take-aways

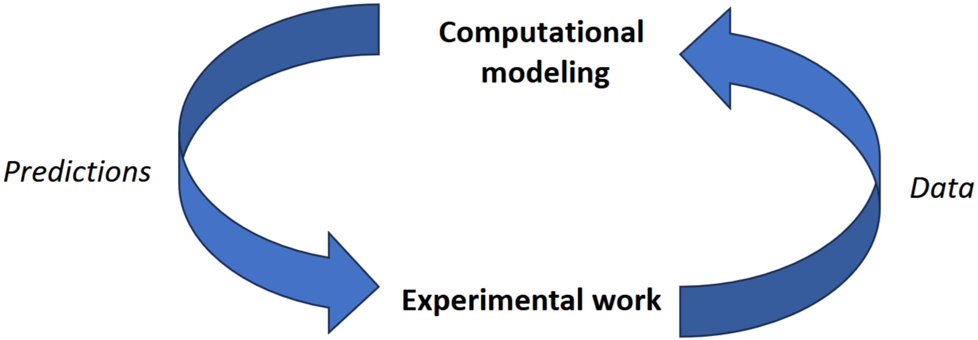
Why computational modeling?

Computational modeling and experimental work constitute a ‘virtuous cycle’



Bruce Hayes

- Computational models provide hypotheses to test
- Experimental work generates data to test hypotheses
- Models/hypotheses are refined based on how well they predict data



Why are computational models good at this?

Two reasons:

1. They require us to be completely explicit in the details of the model and therefore the details of the hypothesis

2. They allow us to link abstract theories to quantitative data

What’s in store

I’ll present two studies that have the same general workflow

1. Deploy models that instantiate different hypotheses on experimental data

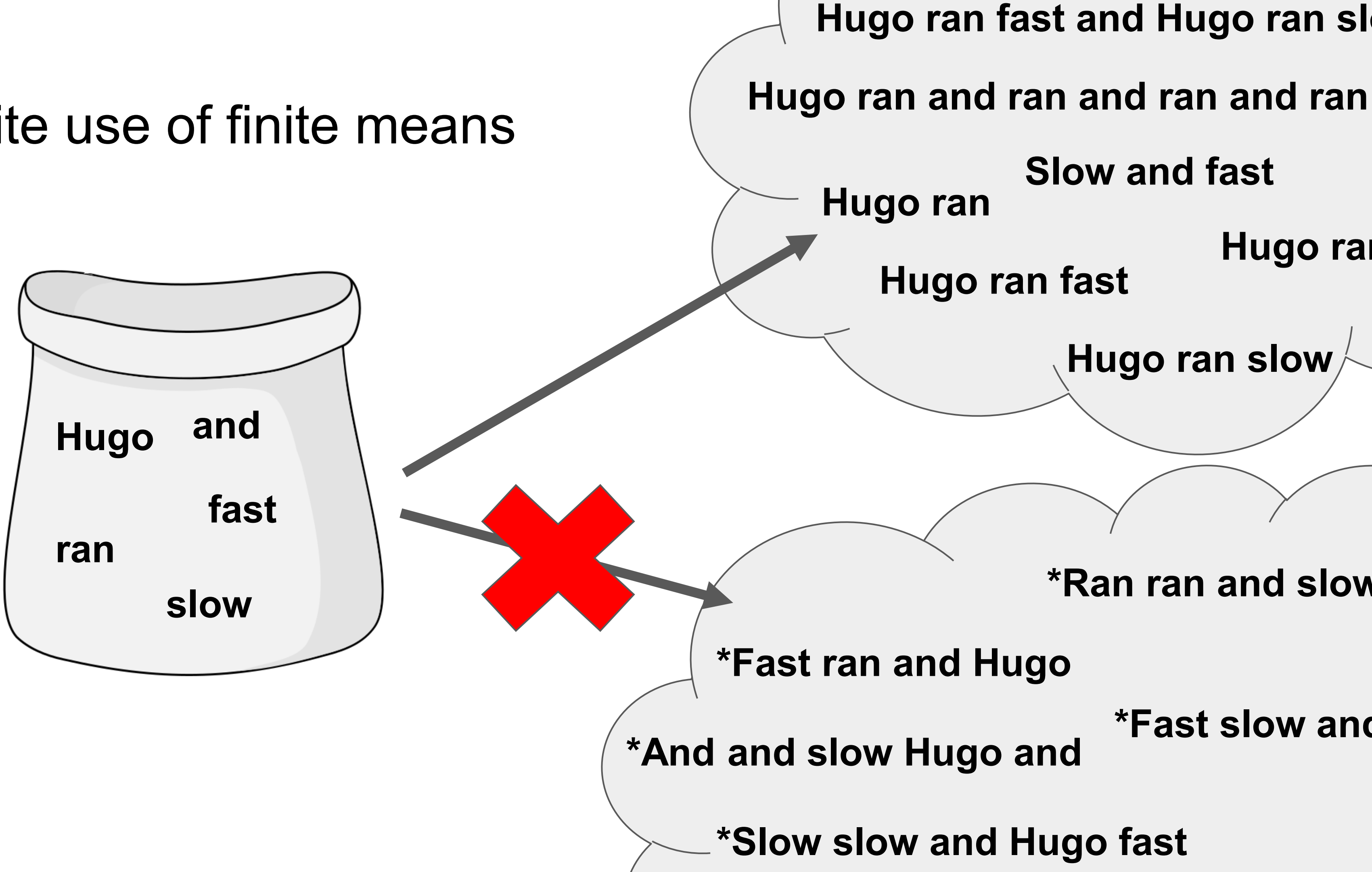
2. Evaluate which models best predict the data

3. Reflect on the properties of each model and (hopefully) learn something

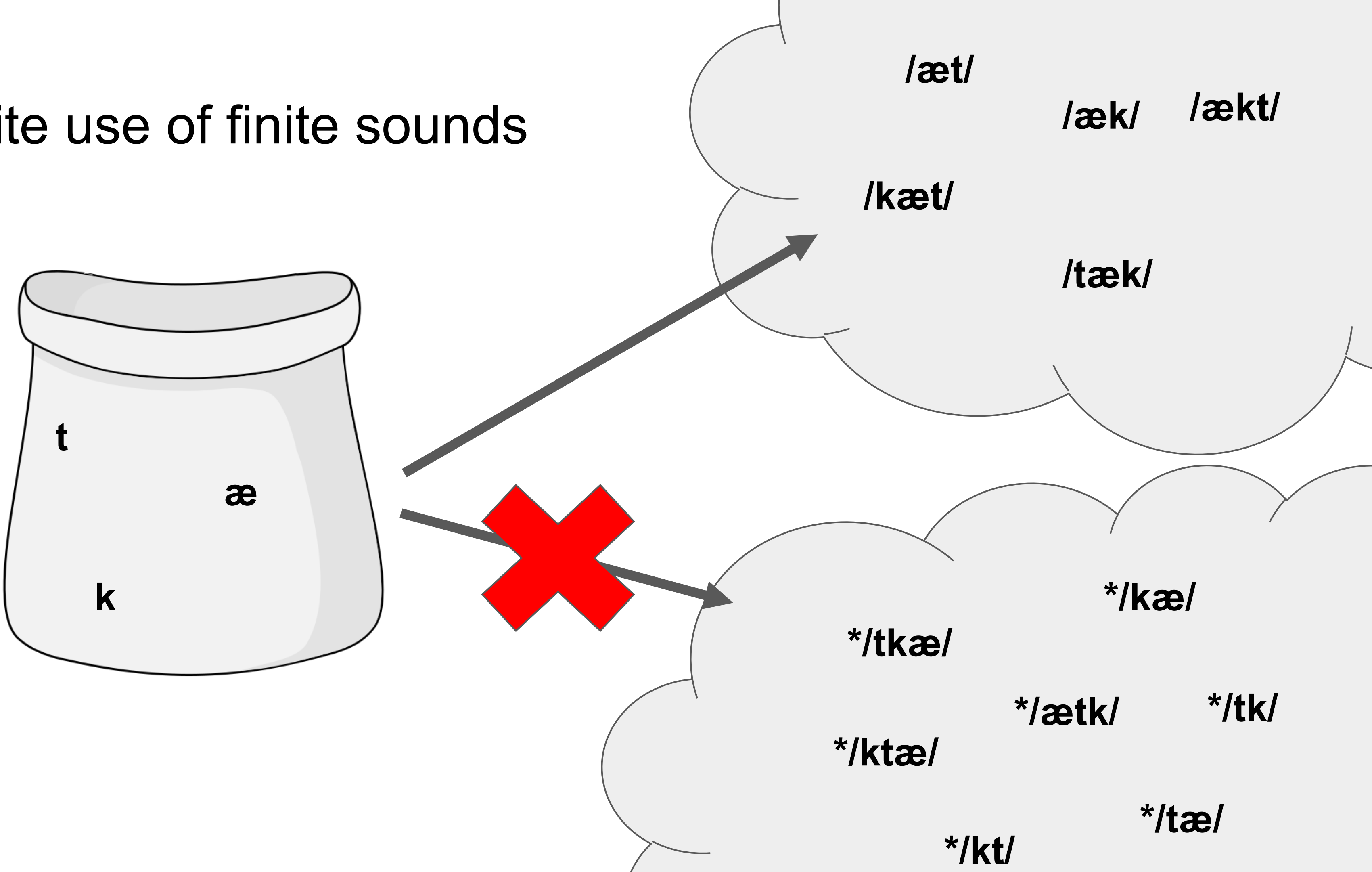
Roadmap

1. Why computational modeling?
2. Background on phonotactics
3. Relating phonotactic learning and word learning
4. Phonotactic model bake-off
5. Discussion and take-aways

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:

- /stik/ would be a fine English word, but not a good Spanish word
- /kwakwəkəʔwakw/ 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 gradience?

poik

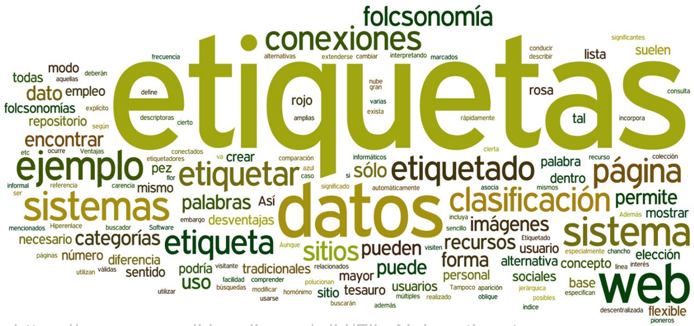
lvag

kip

What do we mean by gradience?

lvag $\lll$ poik $\lll$ kip

Where does phonotactic knowledge come from?



https://commons.wikimedia.org/wiki/File:Nube-etiquetas.png

Lexicon

→ **Generalization based on frequency** →

***#sk**
***#st**
***wu**
***ji**
***ɲ#**
…

Phonotactic knowledge

E.g. Chomsky and Halle (1965, 1968), Bybee (1995, 2003), Pierrehumbert (2001), Bailey & Hahn (2001), Daland et al. (2011), a.o.

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?

Hypotheses

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)

Prelexical hypothesis

Infants learn phonotactics from unparsed utterances

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

Strong lexical hypothesis

Infants learn phonotactics from words they have associated with referents

(Sundara & Breiss, resubmitted)

Support for each perspective

Prelexical hypothesis

- Computationally feasible
- Infants attend to prosodic cues to utterance boundaries (Christophe, Guasti, Nespor, Dupoux & van Ooyen, 1997; Johnson & Seidl, 2008)

Protolexical hypothesis

- 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 CVC word forms that adults were most sensitive to
- Varied in their unigram and bigram probabilities

Phonotactic probabilities

We use the Phonotactic Probability Calculator to quantify phonotactic probability
(Vitevitch & Luce 2004)

- Higher probability → more 'typical' word

Two types of probabilities:

- Unigram: reflects individual segment frequency, not considering order
- Bigram: reflects biphone frequency, sensitive to (local) ordering

Frequency is calculated from a training corpus of word types

Infant experiments (Sundara & Breiss, resubmitted)

Monolingual English learning 5-month-olds

- > 90% exposure to English

Three experiments

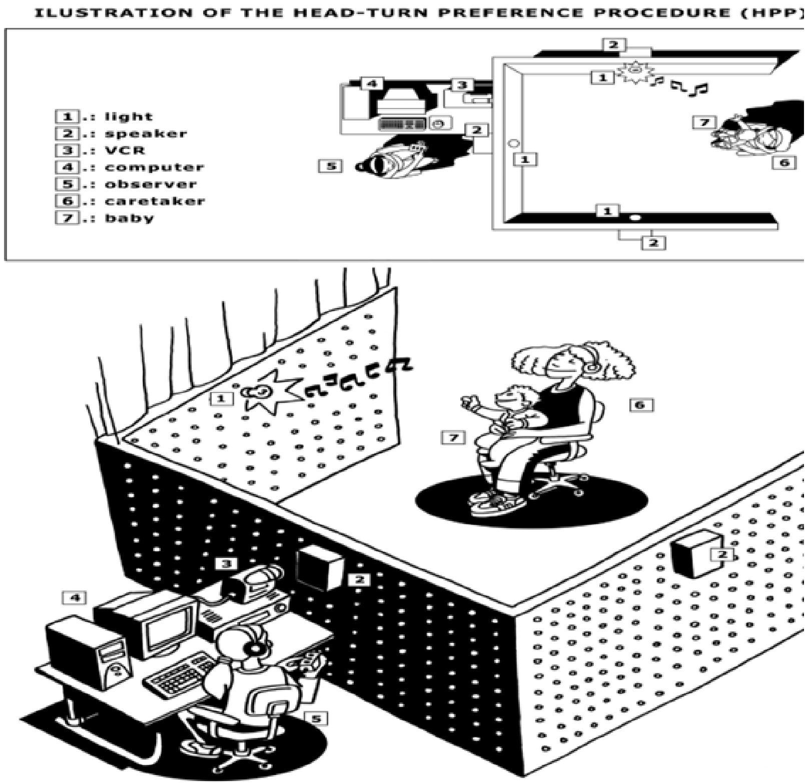
- 2a: High vs. low unigram probability, low bigram probability **(n=30)**
- 2b: (Less) high vs. low unigram probability, low bigram probability **(n=30)**
- 2c: High vs. low unigram and bigram probability **(n=38)**

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



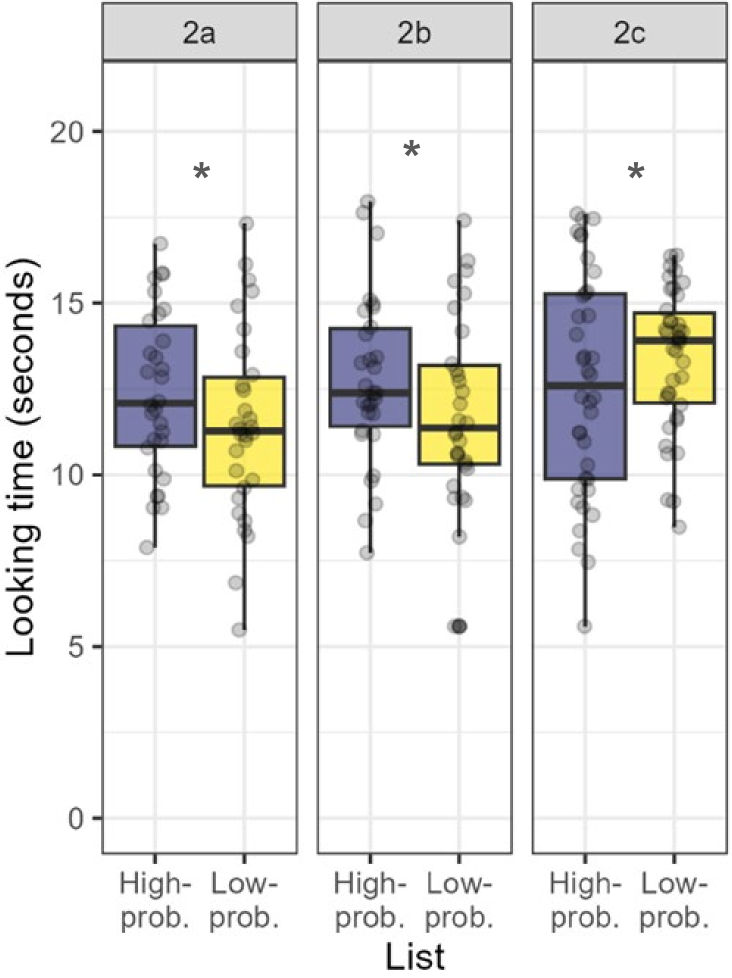
Results

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

Have both cues makes it easier for infants!

- And results in novelty preference
  (Hunter & Ames 1988)

<u>We now have three stimulus sets that 5-mo-olds can distinguish</u>



Study 1: Phonotactics and word learning

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

Modeling phonotactic learning

We want to test the three hypotheses about phonotactic learning

Approach:

1. Create a corpus embodying each hypothesis
2. Calculate unigram and bigram frequencies from corpus
3. Use frequencies to score experimental stimuli for unigram/bigram probability
4. Test if assigned probabilities distinguish high vs. low probability words

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#

#ohnoonewantstogetdressed#

2: Strong lexical hypothesis

Infants learn phonotactics from words with associated referents

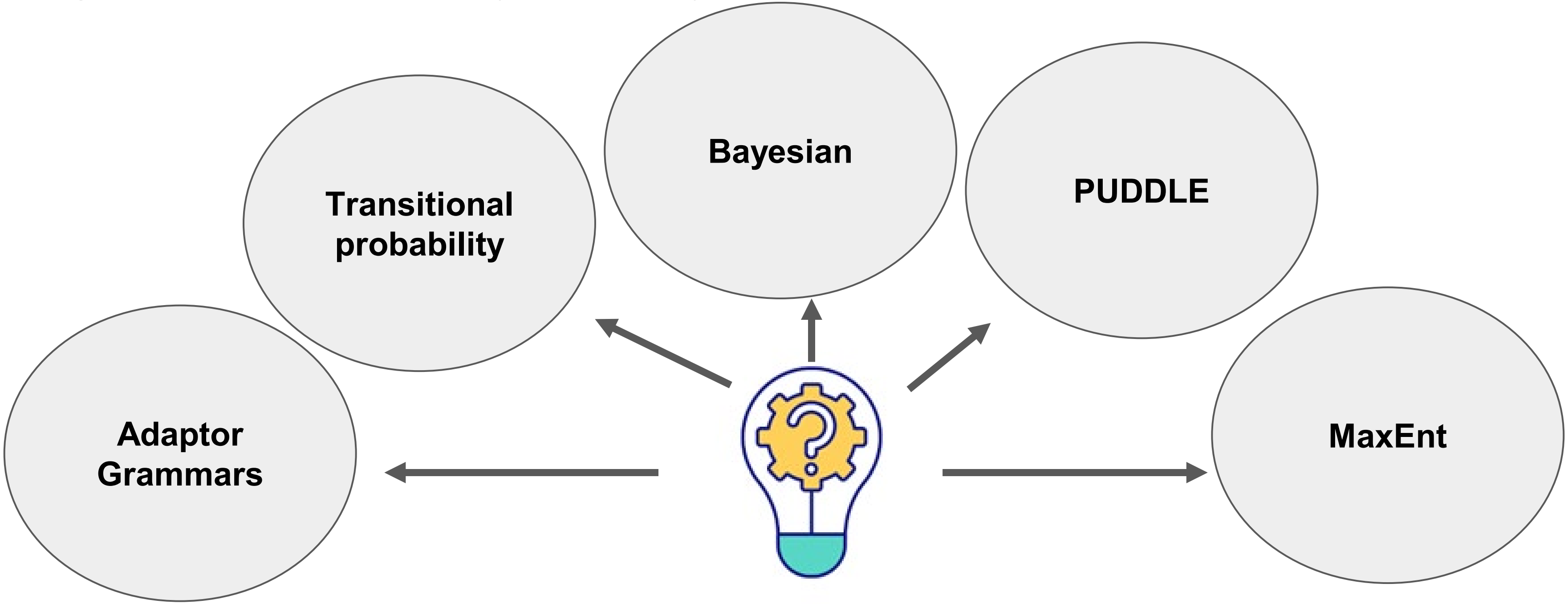
At 5-months, infants associate some word forms with referents
(Bergelson & Swingley, 2012; Bortfeld et al., 2005)

- *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

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



Word segmentation

noitɪŋdɔgfud

Word Segmentation Model

no itɪŋ dɔg fud

(Brent & Siskind, 2001)

Comparing model properties

Model	Joint inference?	Uses stored words for segmentation?	Phonotactics-driven segmentation
MaxEnt	Words and phonotactics	Yes	Yes
Adaptor Grammars	Words and sub/supra-word chunks	Yes	Yes?
PUDDLE	Words and phonotactics	Yes	Yes
Bayesian Unigrams	No	Yes	No
Bayesian Bigrams	Words and preceding word context	Yes	No
Transitional probability	No	No	Yes

3: Protolexical hypothesis

Infants learn phonotactics from word forms in the lexicon

(Thiessen, Kronstein & Huffnagle, 2013)

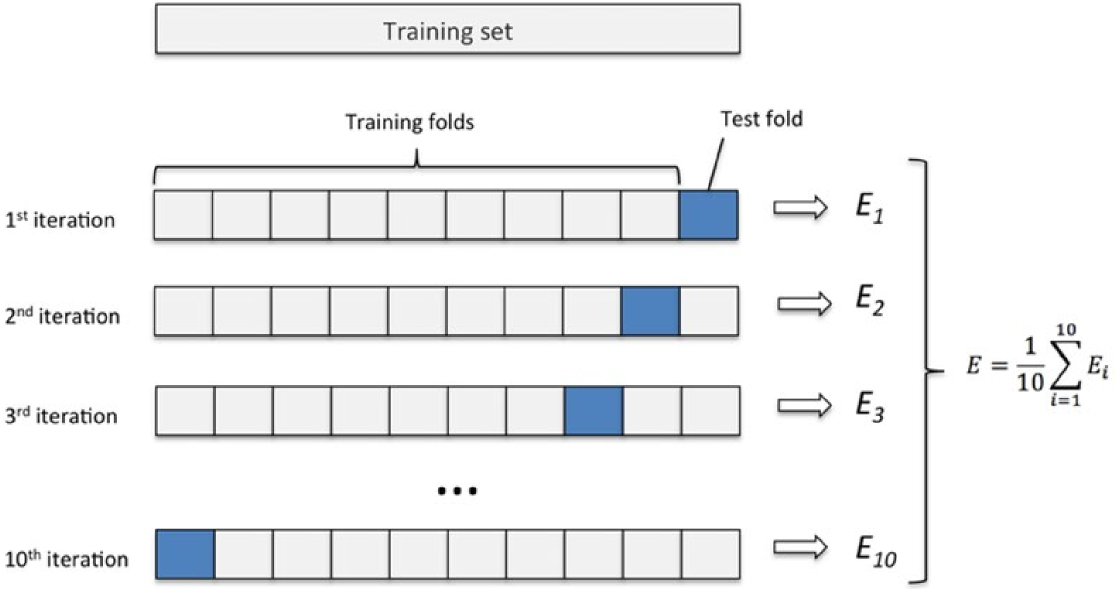
Corpus: Output of 24 unsupervised models of word segmentation on Pearl-Brent corpus of infant-directed speech

- 24 distinct hypotheses about word segmentation strategies

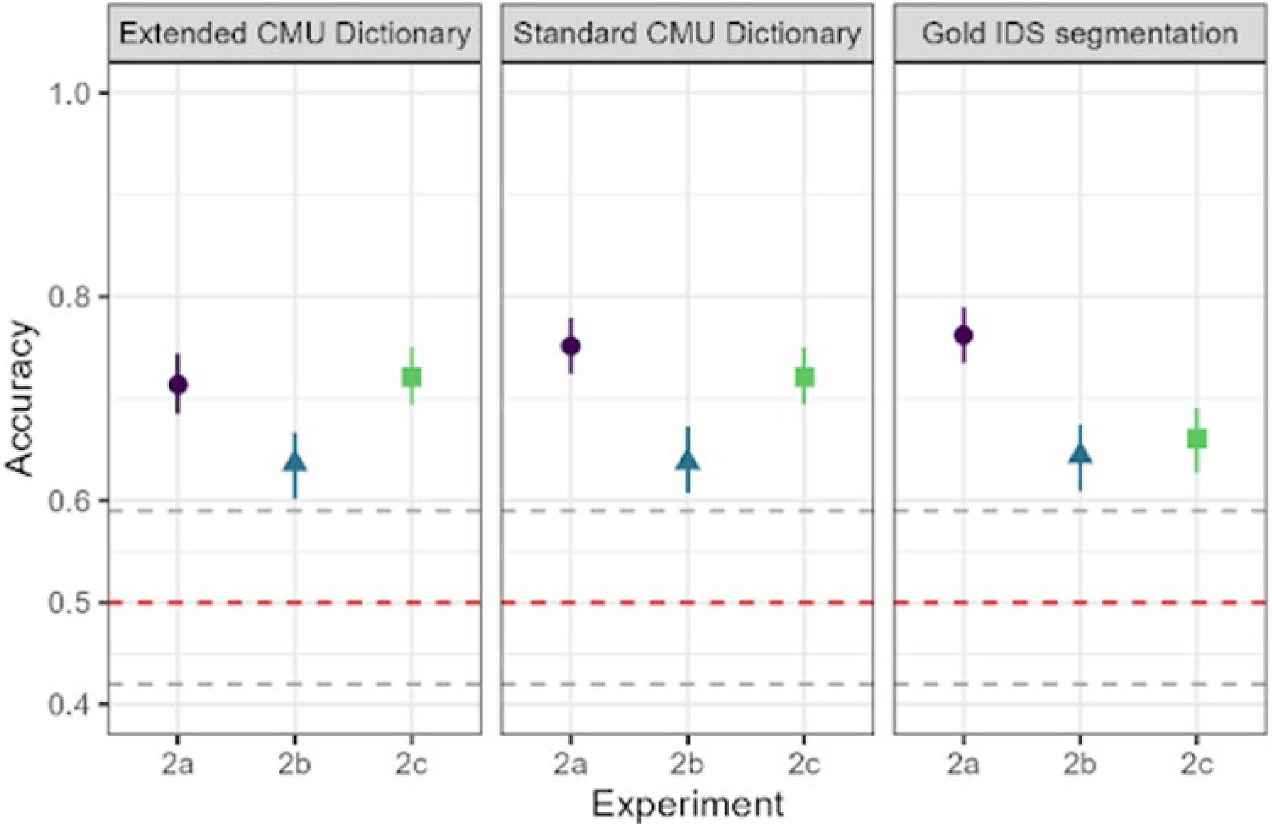
Mostly run using wordseg (Bernard et al. 2019)

Logistic regression model with k-fold cross-validation

High vs. low probability word ~ unigram_probability * bigram_probability

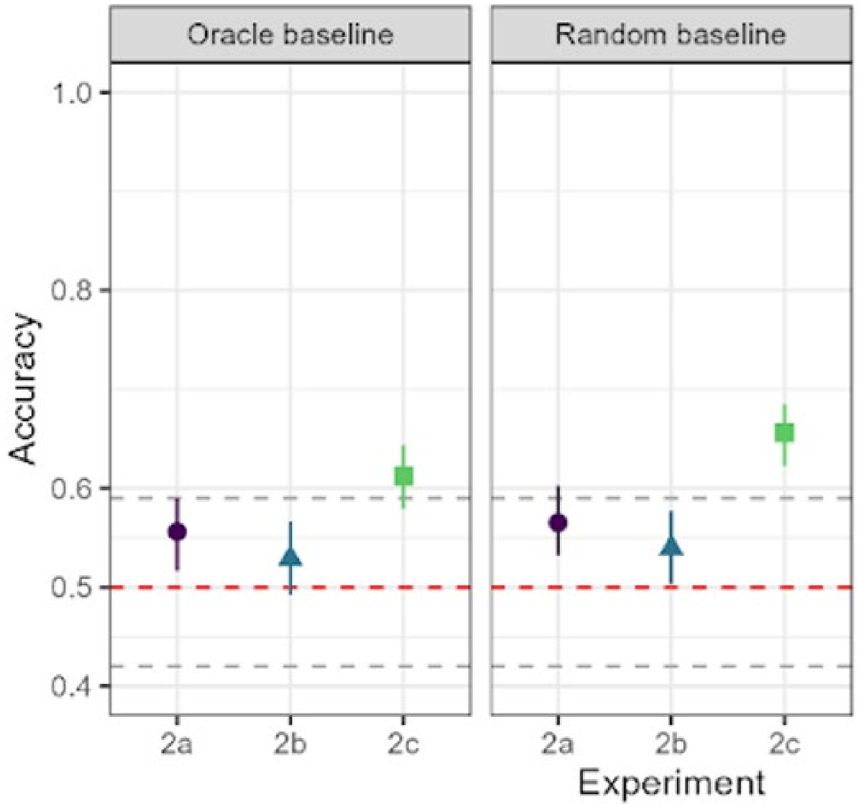


Sanity Check



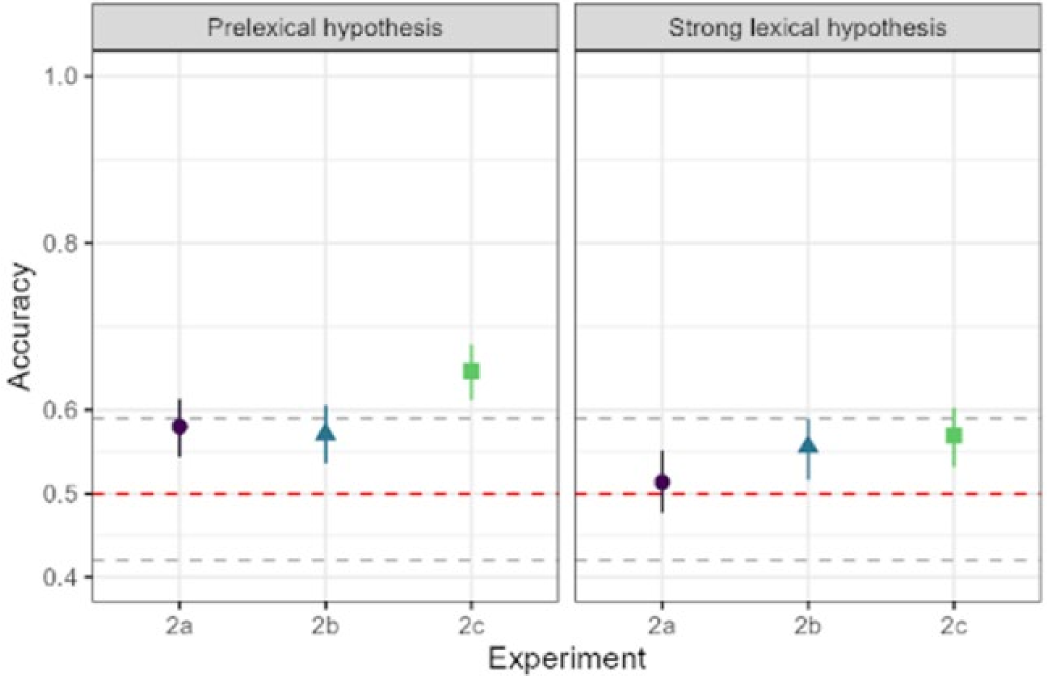
Adult lexicons & fully-segmented infant-directed speech provide sufficient information to distinguish lists distinguished by 5-month-olds.

Baselines



Both baselines provide sufficient information to distinguish list 2c!

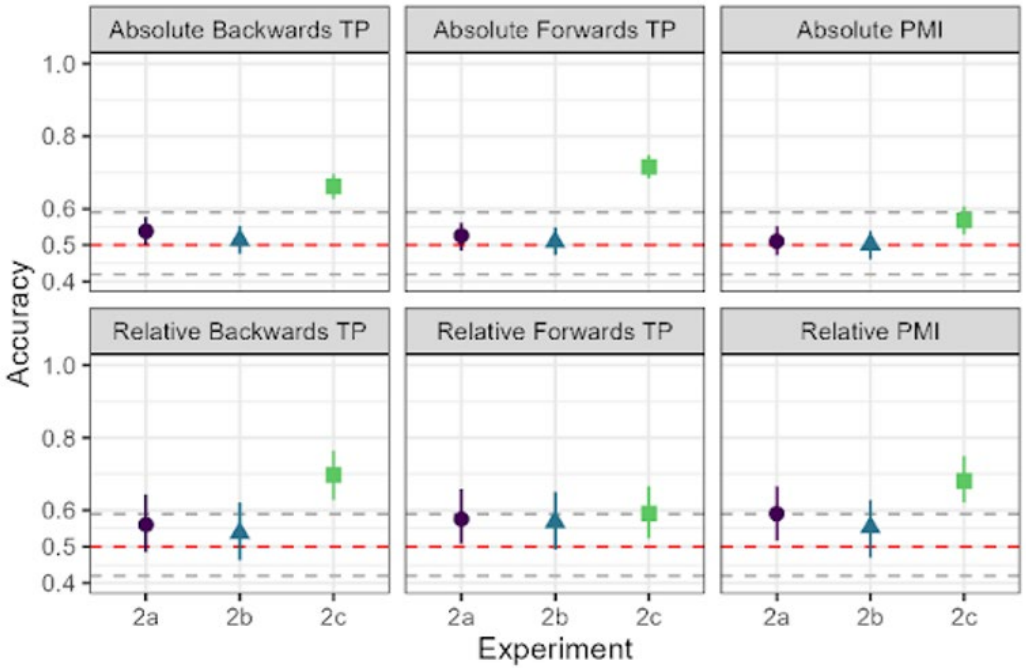
Prelexical and Strong Lexical Hypotheses



Prelexical hypothesis = Baseline

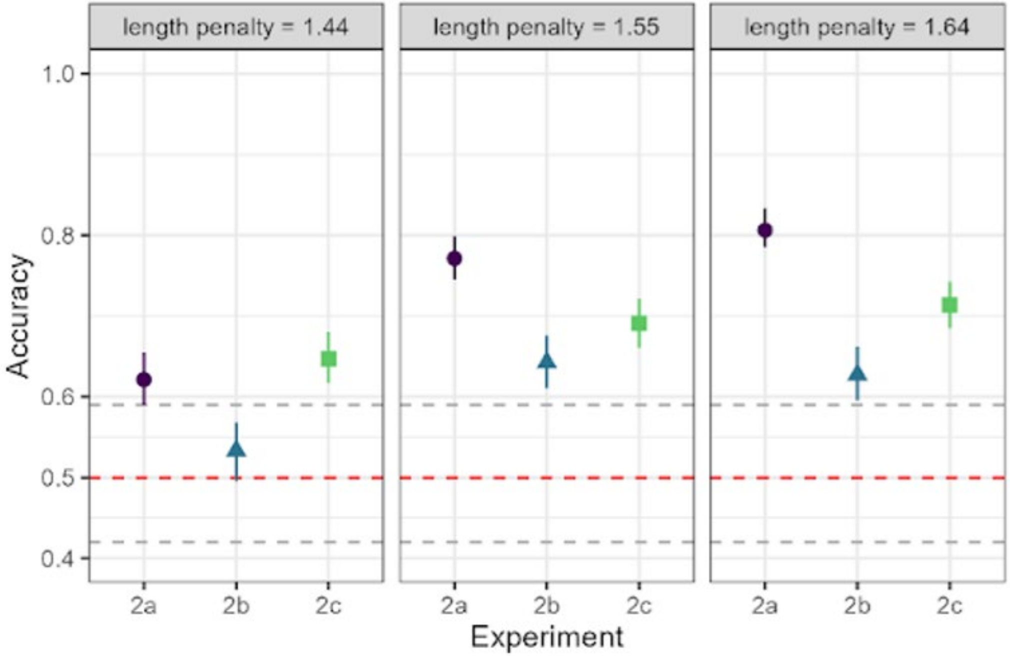
Transitional Probability-based models (Saksida et al. 2017)

Best TP-based model = Baseline



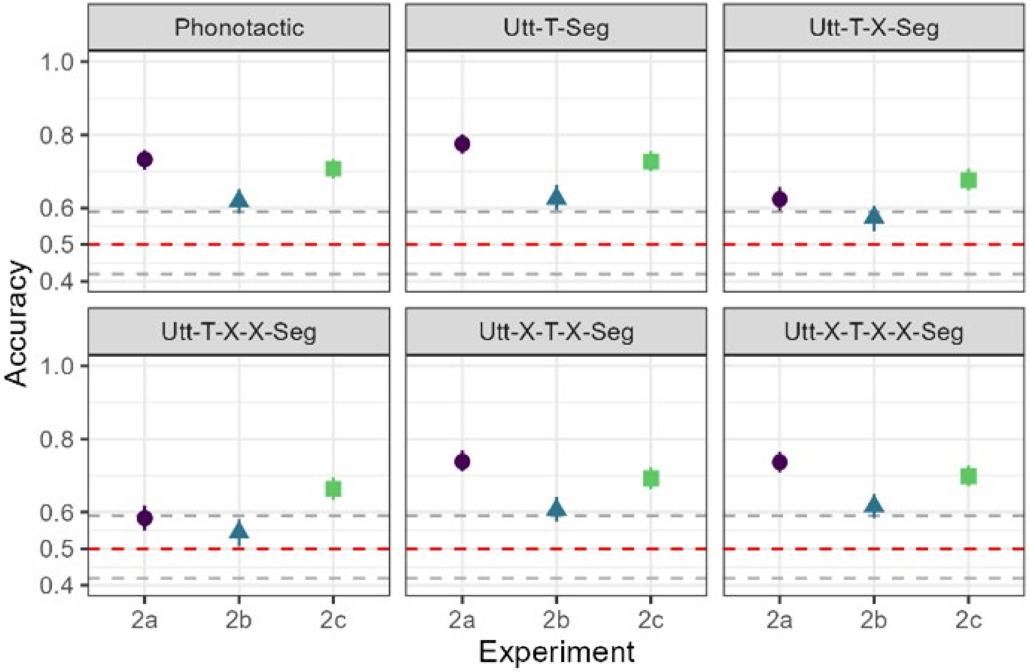
MaxEnt models (Johnson, Pater, Staubs & Dupoux, 2015)

Two of three models distinguish all lists



Adaptor grammar models (Johnson et al. 2006)

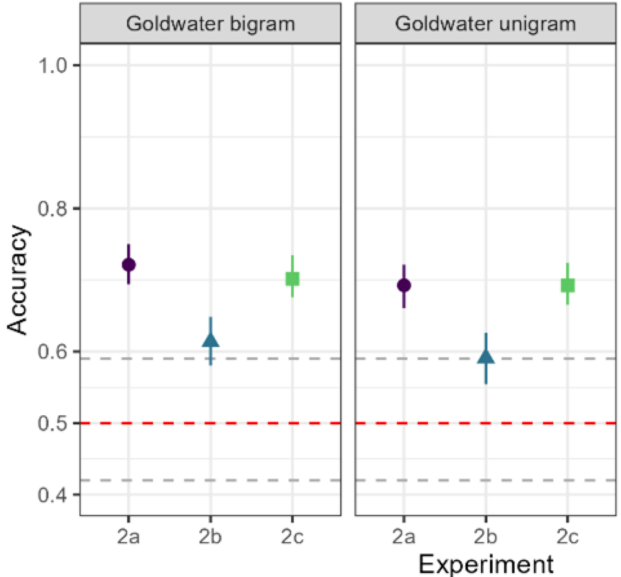
Four of six models distinguish all three lists



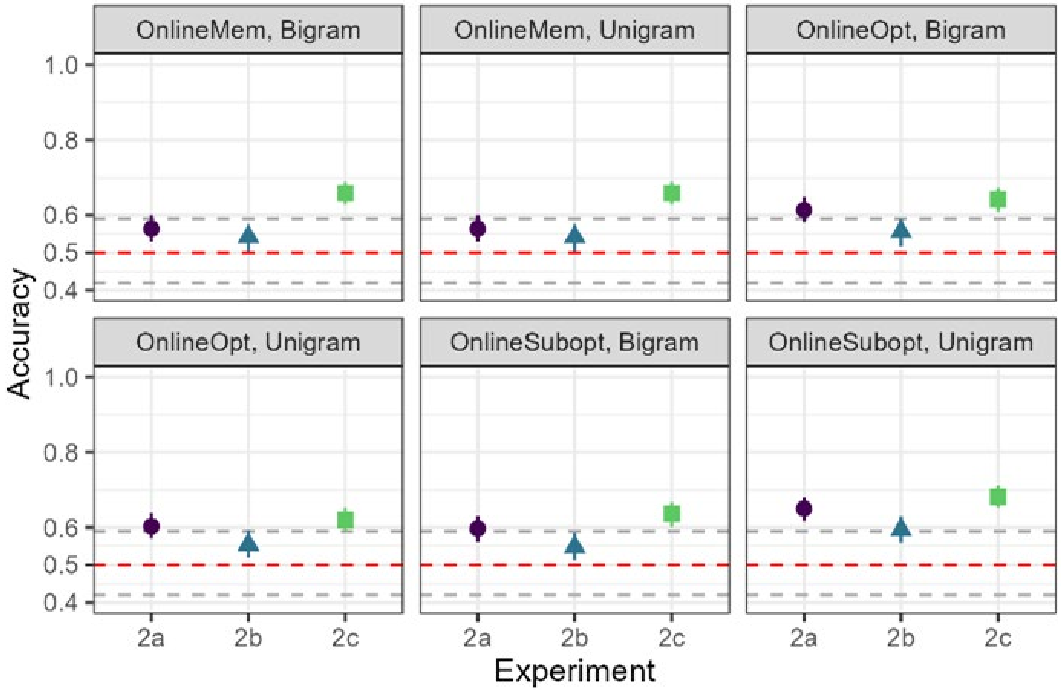
Bayesian models

One of eight models distinguishes all three lists

Goldwater et al. (2009)

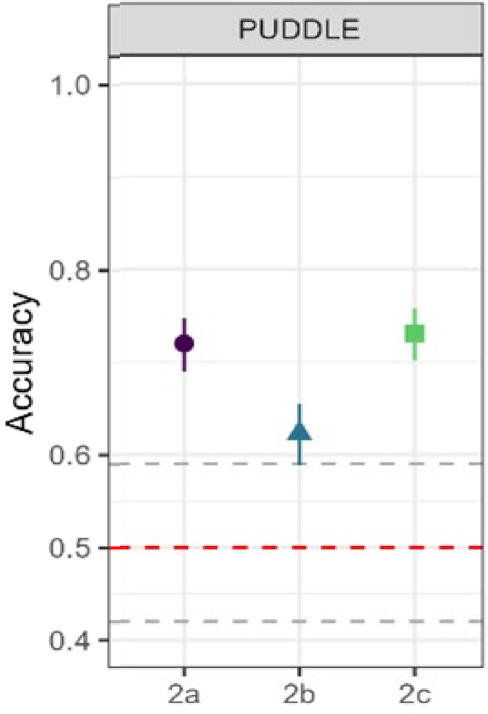


Phillips & Pearl (2015)



PUDDLE (Monaghan and Christiansen 2010)

PUDDLE distinguishes all three lists



Summary: Protolexical Hypothesis

11 of 24 models do no better than baselines

- All TP-based models (Saksida et al. 2017)
- Cognitively plausible Bayesian models (Phillips & Pearl 2015)
- One adaptor grammar model (Johnson, Griffiths & Goldwater 2006)

Only **8 of 24** distinguished items in all three lists

- Adaptor grammar models (4 of 6; Johnson, Griffiths & Goldwater 2006)
- MaxEnt models (2 of 3; Johnson, Pater, Staubs & Dupoux 2015)
- Bigram Bayesian learning model (Goldwater, Griffiths & Johnson 2009)
- PUDDLE (Monaghan & Christiansen 2010)

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!

Comparing model properties

Model	**Joint inference?**	**Uses stored words for segmentation?**	**Phonotactics-driven segmentation**
MaxEnt	**Words and phonotactics**	**Yes**	**Yes**
Adaptor Grammars	**Words and sub/supra-word chunks**	**Yes**	**Yes?**
PUDDLE	**Words and phonotactics**	**Yes**	**Yes**
Bayesian Unigrams	No	Yes	No
Bayesian Bigrams	*Words and preceding word context*	*Yes*	*No*
Transitional probability	No	No	Yes

Evaluating mechanisms

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

- Prelexical hypothesis ✗
- Strong lexical hypothesis ✗
- Protolexical hypothesis ✓ **(some proposals)**

Successful protolexical models use **joint learning**, rely on **stored words** to bootstrap segmentation, and apply **phonotactic restrictions** to segmentation.

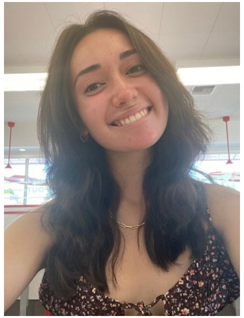
Caveat: All protolexical hypotheses are better at segmenting words than 5-month-olds!

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

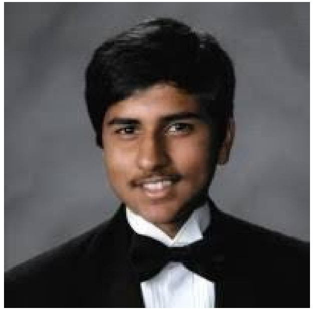
Roadmap

1. Why computational modeling?
2. Background on phonotactics
3. Relating phonotactic learning and word learning
4. Phonotactic model bake-off
5. Discussion and take-aways

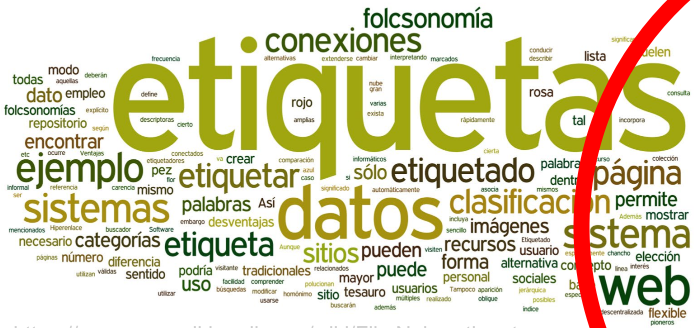
Study 2: Comparing models of phonotactics

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

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



Where does phonotactic knowledge come from?



https://commons.wikimedia.org/wiki/File:Nube-etiquetas.png

Lexicon

Generalization based on frequency

***#sk**
***#st**
***wu**
***ji**
***ŋ#**
…

Phonotactic knowledge

E.g. Chomsky and Halle (1965, 1968), Bybee (1995, 2003), Pierrehumbert (2001), Bailey & Hahn (2001), Daland et al. (2011), a.o.

Modeling phonotactic knowledge

Goal: we want a computational model that reflects human phonotactic knowledge

- Model should score words in a way that tracks with human behavior

All the models we consider treat phonotactics as probabilistic

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

Output: How probable is a word $\mathbf{w}$ composed of the segments $\mathbf{x}_1 \ldots \mathbf{x}_n$?

What are we doing here?

We’ll compare two simple and popular models of phonotactic probability based on how well they predict results from acceptability judgment studies.

The models we’ll look at will include

- A venerable model (Markov 1913, Shannon 1948)
- A more recent proposal (Vitevitch and Luce 2004)

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.

Quantifying phonotactic probability

Different models have been applied to quantify phonotactic probability

- N-gram models (Markov 1913, Shannon 1948, Vitevitch and Luce 2004, Albright 2009)
- Maximum Entropy models (Hayes & Wilson 2008, Dai, Mayer and Futrell 2024)
- Neural networks (Mirea and Bicknell 2019, Mayer and Nelson 2020)

And different representational assumptions

- Segmental (Shannon 1948, Vitevitch and Luce 2004)
- Subsegmental (everything else above)

Quantifying phonotactic probability

Different models have been applied to quantify phonotactic probability

- **N-gram models** (Markov 1912, Shannon 1948, Vitevitch and Luce 2004, Albright 2009)
- Maximum Entropy models (Hayes & Wilson 2008)
- Neural networks (Mirea and Bicknell 2019, Mayer and Nelson 2020)

And different representational assumptions

- **Segmental** (Shannon 1948, Vitevitch and Luce 2004)
- Subsegmental (everything else above)

Why segmental n-grams?

They’re still **widely used in research** contexts

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

They’re **simple** to implement and reason about

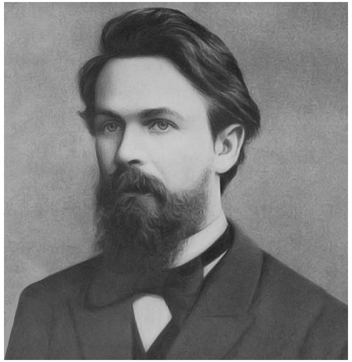
They **get us reasonably far in phonotactics**

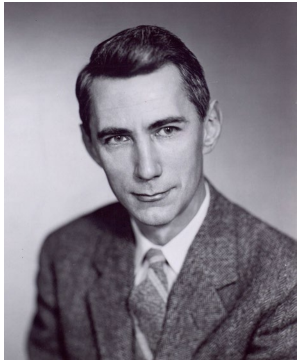
- Bigram model on English onset acceptability judgment data ***r = 0.877***
  (Daland et al. 2011, Dai, Mayer and Futrell 2023)

Two prominent n-gram models

Researchers often use one of two n-gram models

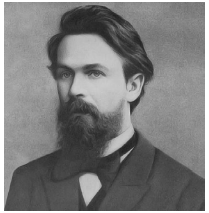
1. **Standard n-grams** (Markov 1913, Shannon 1948)

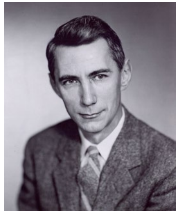




2. **Phonotactic Probability Calculator** (Vitevitch and Luce 2004)

The standard n-gram model

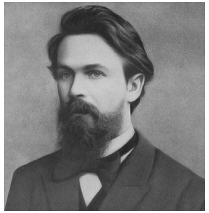
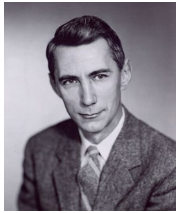




Unigram model: $$P(w = x_1 \ldots x_n) \approx \prod_{i=1}^{n} P(x_i)$$

Bigram model: $$P(w = x_1 \ldots x_n) \approx \prod_{i=2}^{n} P(x_i | x_{i-1})$$

Estimating probabilities from data

We can estimate probabilities by counting occurrences in a corpus

$$P(x) = \frac{C(x)}{\sum_{y \in \Sigma} C(y)}$$

Unigrams: Of the times I see a segment, in what proportion is it `x`

$$P(x|y) = \frac{C(yx)}{C(y)}$$

Bigrams: Of the times I see `y`, in what proportion is the following segment `x`

Padding

In standard n-gram models, boundary symbols are inserted at word edges

`/skif/` → **`/#skif#/`**

Allows bigrams to refer to word boundaries

- **`P(s|#)`** – the probability that a word begins with **`s`**

The Phonotactic Probability Calculator

$$PosUniScore(w = x_1 \ldots x_n) = 1 + \sum_{i=1}^{n} P(w_i = x_i)$$

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

where $\mathbf{w}_{\mathbf{i}}$ is the segment in the $\mathbf{i}^{th}$ position in word $\mathbf{w}$

Major difference 1: The PPC considers *absolute position* within the word

Major difference 2: The PPC combines probabilities using *addition*

Estimating probabilities from data in the PPC

$$P(w_i = x) = \frac{C(w_i = x)}{\sum_{y \in \Sigma} C(w_i = y)}$$

Unigrams: Of the times I see a segment *in position `i`*, in what proportion is it **`x`**

$$P(w_{i-1} = y, w_i = x) = \frac{C(w_{i-1} = y, w_i = x)}{\sum_{z \in \Sigma} \sum_{v \in \Sigma} C(w_{i-1} = z, w_i = v)}$$

Bigrams: Of the times I see a pair of segments *in positions `i-1` and `i`, in what proportion is that pair* **`yx`**

Major difference 3: The PPC uses *joint probabilities*

Other details about the PPC

<u>Major difference 4</u>: The PPC does not use word boundary symbols

Position 1 always corresponds to word-initial position

Word-final position cannot be represented in the model

- Position 3 is word-final in [dɔ**g**] but not in [it**ɪ**ŋ]

Summary of model differences

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

A comment on phonological theory

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

Hayes (2012) notes that phonologists would it "extremely controversial"

Phonologies don't count large numbers (McCarthy & Prince 1986)

- Ideas like "the 7th segment in the word" don't seem to be helpful
- When counting happens, it's usually related to prosodic structures, not segments

Many phonotactic restrictions are related to word-final position!


THE GREAT
PHONOTACTIC
BAKE OFF

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

Let’s compare the standard n-gram and PPC 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
Albright & Hayes (2003)	English	20	58 3-5 segment, monosyllabic nonce verbs	Likert scale	Auditory
Daland et al. (2011)	English	48	96 disyllabic nonce words differing in the initial onset	Likert scale	Orthographic
Needle et al. (2022)	English	1440	8400 nonce words, between 4-7 segments	Likert scale	Orthographic
Scholes (1966)	English	33	62 monosyllabic nonce words differing in initial onset	Forced choice	Orthographic
Hayes & White (2013)	English	29	160 nonce words, between 2 and 7 segments	Magnitude estimation	Orthographic and auditory

Datasets

Paper	Lang	Subjects	Stimuli	Input	Presentation
Jarosz & Rysling (2017)	Polish	81	159 nonce words varying in onset properties	Likert scale	Orthographic
Mayer & Sundara (in prep)	Spanish	168	575 CVC nonce words	Magnitude estimation	Orthographic and auditory
Mayer (in press)	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. Score each of the test stimuli using the trained models

3. Predict participant responses with a (linear/logistic) regression model

```
response ~ uni_prob * bi_prob
```

4. Compare models using AIC (Akaike 1974)

AIC Rules of Thumb

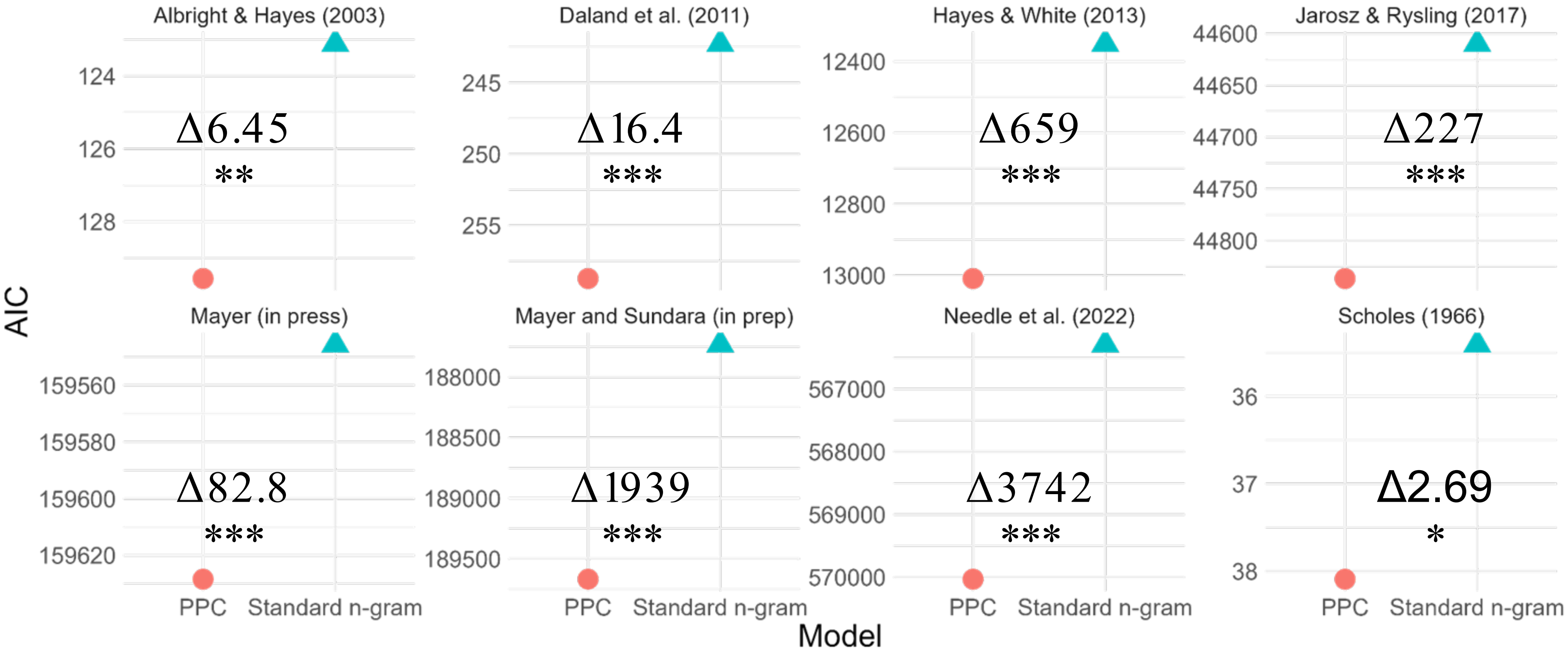
AIC is an estimate of prediction accuracy on held-out data

- We interpret AIC in terms of <u>differences</u> between models
- <u>*Lower AIC indicates better fit to data*</u>

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

- **ΔAIC ≤ 2**: no difference between models
- **ΔAIC > 10**: strong support for model with lower AIC
- Increasing **ΔAIC** indicates increasing certainty in better model

Standard n-grams are better in every case



Model Bake-off 2: but *why?* (Mayer & Sundara in prep)

The two models differ on four dimensions

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

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

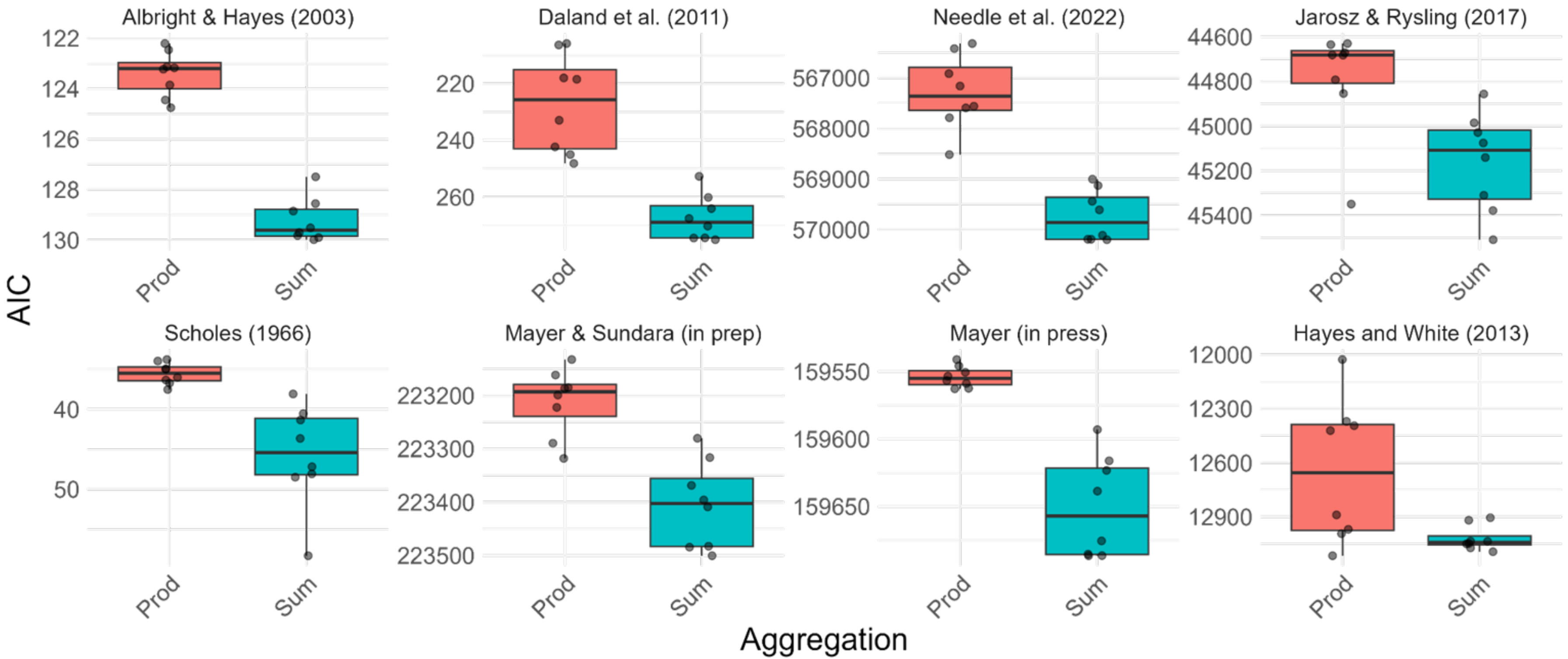
Bake-off 2 procedure

We implemented *16 different models for each combination* of these parameters

- One model per possible combination of the four parameters

We fit each model to each dataset

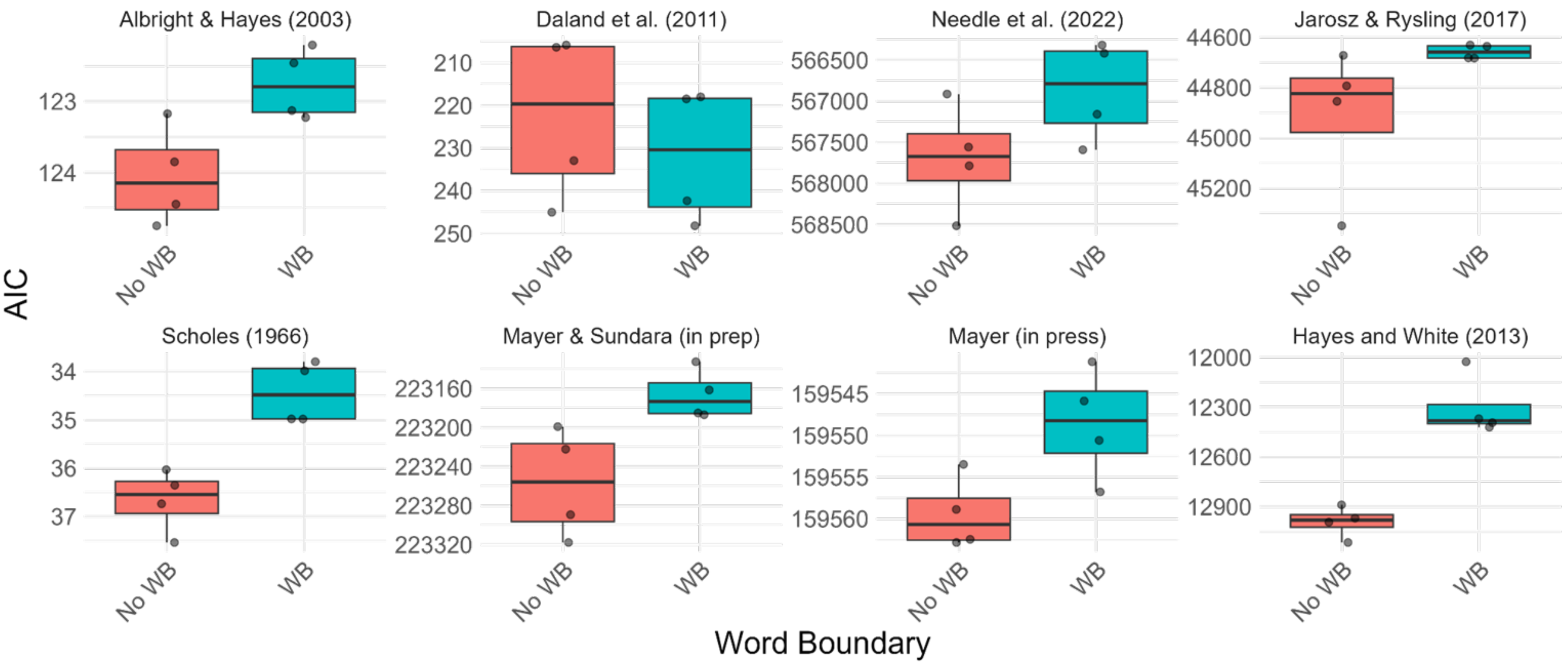
Result 1: Adding probabilities is almost always worse



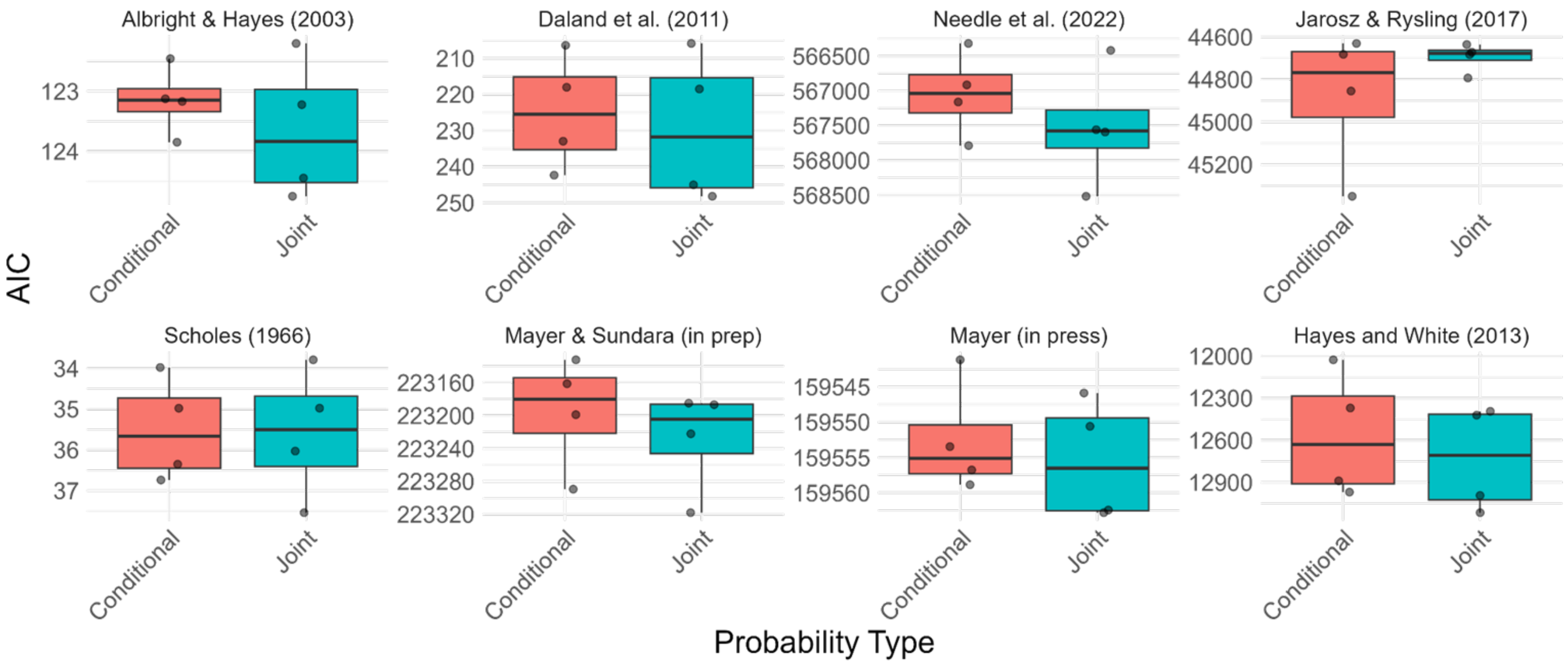
We’ll only consider the ‘product’ models going forward

$$\Pi$$

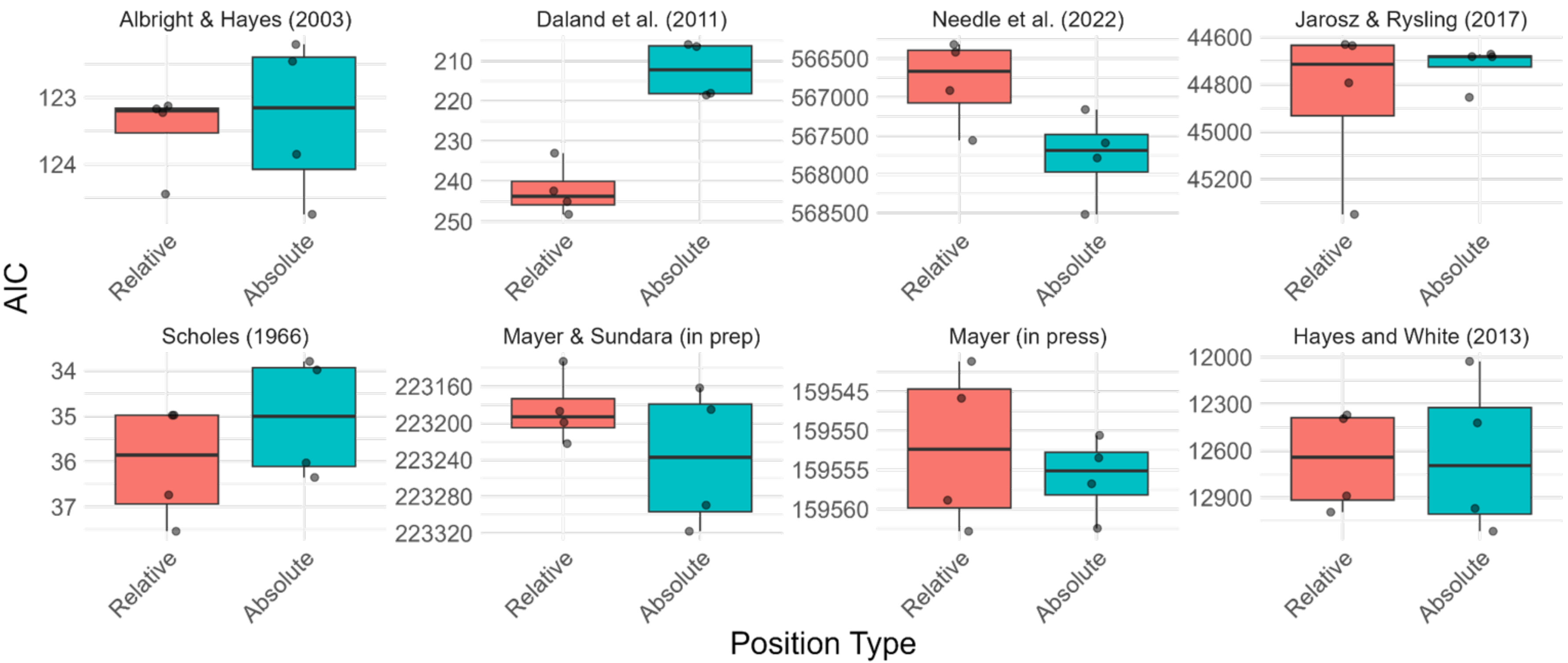
Result 2: Word boundaries help



Result 3: A weak preference for conditional probabilities



Result 4: Relative vs. absolute varies across dataset



Bake-off 2 Results

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

What makes a good phonotactic model?

Immediate practical consequence

PPC is less predictive of acceptability judgments than standard n-gram models across all the data sets we examined

Theoretical perspective

We can say something about desiderata for a phonotactic models

Zooming in on model properties

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 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. <u>**Conditional probabilities > joint probabilities**</u>
    - The two are highly correlated (Gaygen 1997, Vitevitch and Luce 1999)
    - Only conditional probabilities get us a valid probability distribution

4. <u>**Absolute vs. relative position varied across datasets**</u>
    - General preference for relative
    - Likely related to specific data sets used
        1. bigrams can’t ‘see’ full #CC onsets in Daland et al. (2011)
        2. Positional model can track (some of) this information

More support for relative position

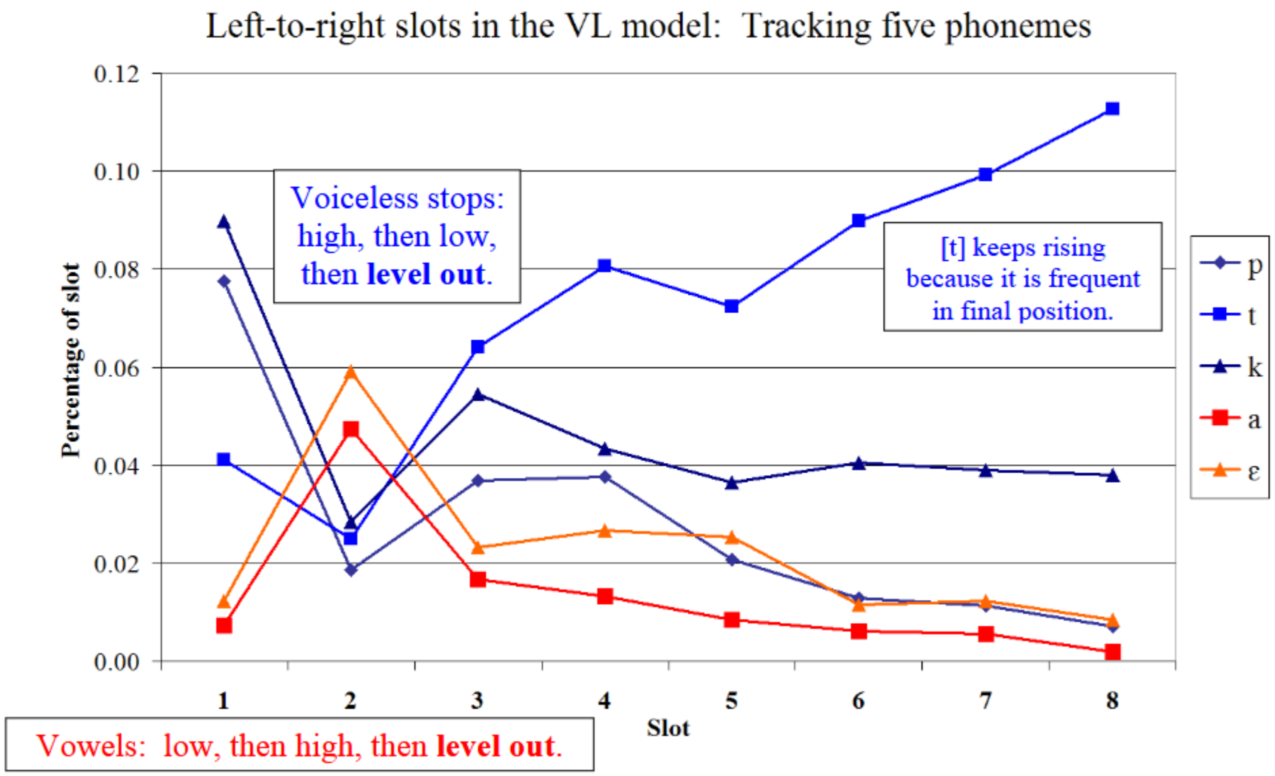
The mean length of an English word is about 6 segments (Marian et al. 2012)

- Lots of data for positions 1, 2, 3, …
- Less data for positions 10, 11, 12..

We run into data scarcity issues as words get longer

- Words in these 8 studies are mostly short, often the same length/template
- Needle et al. (2022) has the greatest variability in word length
- It is also one of the datasets that most strongly favors relative position models

A plot from Hayes (2012)



Limitations and next steps

Phonotactics is relevant to other downstream tasks:

- 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)

Are the best metrics for acceptability judgments the best in these domains?

(cf. Castro and Vitevitch 2023)

Roadmap

1. Why computational modeling?
2. Background on phonotactics
3. Relating phonotactic learning and word learning
4. Phonotactic model bake-off
5. 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

Comparing the predictions of computational models against experimental data allows us to make some claims about how phonotactic learning must progress.

Study 1: 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:

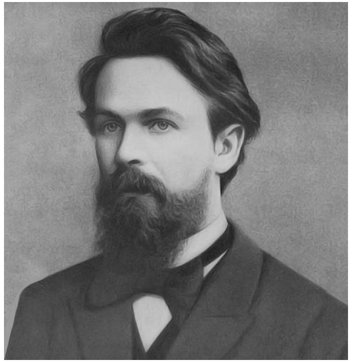
1. They employ joint learning (words + something else)
2. They use previously identified words to bootstrap segmentation
3. They evaluate possible new words based on identified phonotactic restrictions

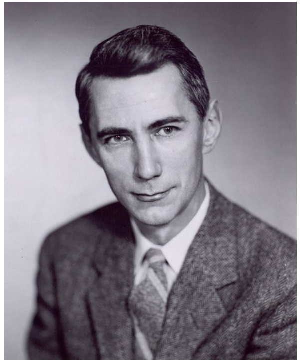
Study 2: Comparing phonotactic models

The **standard n-gram model** most consistently predicts experimental responses

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

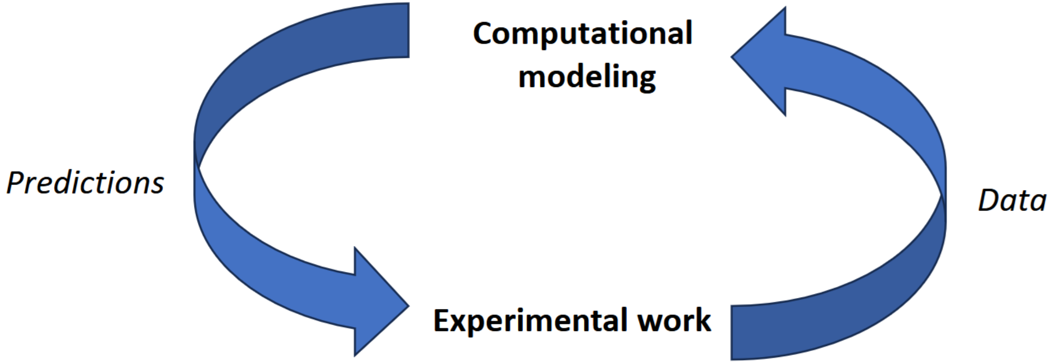
- More complex models will likely need to preserve these useful properties





Closing the loop

Broader goal: “Close the loop” between computational and experimental work



Sharing is caring

“No data ever lose their usefulness”

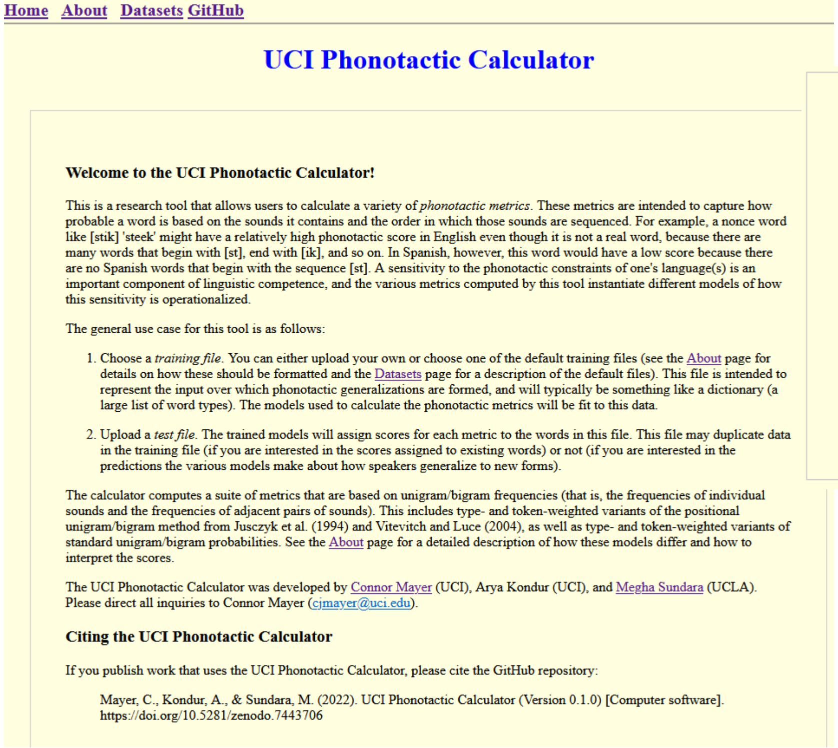
- Hayes (2012)

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

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

- I encourage you all to do the same!

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



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).

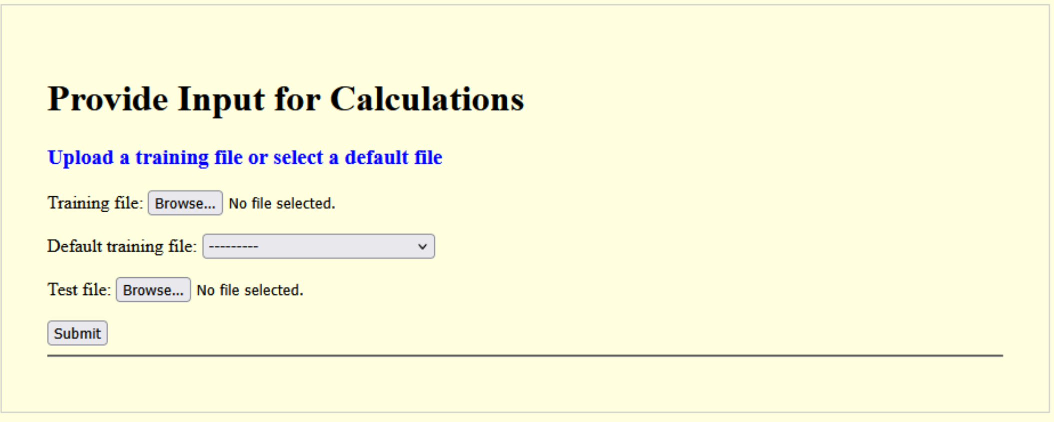
The calculator computes a suite of metrics that are based on unigram/bigram frequencies (that is, the frequencies of individual sounds and the frequencies of adjacent pairs of sounds). This includes type- and token-weighted variants of the positional unigram/bigram method from Jusczyk et al. (1994) and 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).

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: Browse... No file selected.

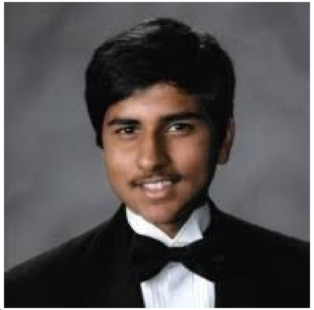
Default training file: ---------

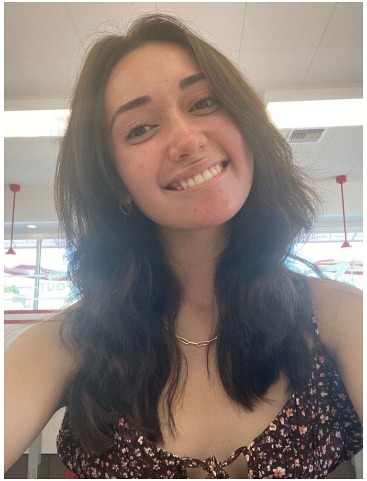
Test file: Browse... No file selected.

Submit

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'). Izvistia 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. Glot 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)

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).

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: Browse... No file selected.

Default training file: ---------

Test file: Browse... No file selected.

Submit

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	AH B AE K	59
3	AE B AH K AH S	8
4	AH B AE N D AH N	1010
5	AH B AE SH	15
6	AH B EY T	42
7	AE B IY	7
8	AE B EY	7
9	AE B IY	181
10	AE B AH T	43
11	AH B R IY V IY EY T	35
12	AH B R IY V IY EY SH AH N	14
13	AE B D AH K EY T	40
14	AE B D IH K EY SH AH N	34
15	AE B D OW M AH N	57
16	AE B D AH M AH N	57
17	AE B D AA M AH N AH L	63
18	AH B D AA M AH N AH L	63
19	AE B D AH K T	19
20	AE B D AH K SH AH N	5.5
21	AH B D AH K SH AH N	5.5
22	AH B EH D	4
23	AE B EH R AH N T	11
24	AE B ER EY SH AH N	50
25	AH B EH T	33
26	AH B EY AH N S	17
27	AE B HH AO R	39
28	AH B HH AO R AH N S	7
29	AE B HH AO R AH N T	23
30	AH B AY D	84
31	AH B IH L AH T IY	1557
32	AE B JH EH K T	57
33	AH B L EY Z	29
34	EY B AH L	5887
35	AE B N AO R M AH L	105
36	AE B N AO R M AE L AH T IY	39
37	AA B OW	6
38	AH B AO R D	285
39	AH B OW D	31
40	AH B AA L IH SH	301

Scored test file

1	word	word_len	uni_prob	uni_prob_freq_weighted	uni_prob_smoothed	uni_prob_freq_weighted_smoothed
2	B L IY G IH F	6	-21.28560225	-21.28547321	-21.36475687	-21.36471595
3	B L EH Z IH G	6	-21.89701032	-21.89653607	-21.96285725	-21.96272277
4	B R IY G IH F	6	-21.26431799	-21.26419239	-21.31293023	-21.31289144
5	B R EH P IH D	6	-19.85093399	-19.85144946	-19.78328505	-19.78342243
6	B W IY G IH F	6	-23.46505863	-23.46365267	-23.44272982	-23.44239313
7	B W AA S IH P	6	-21.82616145	-21.82539077	-21.76996196	-21.76979186
8	D G EH P IH D	6	-20.91194316	-20.91206033	-20.85977901	-20.85980997
9	D G AA T IH F	6	-21.1446086	-21.14449317	-21.17316921	-21.17313346
10	D N IY G IH F	6	-20.55196506	-20.55203056	-20.5815925	-20.58160607
11	D N AA T IH F	6	-19.37124649	-19.37172047	-19.36533752	-19.36546055
12	D R IY G IH F	6	-20.8320401	-20.83206664	-20.83634114	-20.83634568
13	D R EH P IH D	6	-19.4186561	-19.41932371	-19.30669597	-19.30687668
14	D W EH Z IH G	6	-23.64418881	-23.64258978	-23.56424112	-23.5638542
15	D W AA T IH F	6	-21.85206218	-21.85121683	-21.74988576	-21.74970185
16	F L EH Z IH G	6	-22.0996585	-22.09908688	-22.12310698	-22.12295264
17	F L AA T IH F	6	-20.30753186	-20.30771393	-20.30875163	-20.30880029
18	F N IY B IH D	6	-20.05862089	-20.05896282	-20.07368218	-20.07377139
19	F N EH Z IH G	6	-21.7982992	-21.79776998	-21.81653169	-21.81638852
20	F R EH P IH D	6	-20.05358216	-20.05400026	-19.94353478	-19.9436523
21	F R AA S IH P	6	-19.82806898	-19.82848129	-19.80041209	-19.80052004
22	F W IY B IH D	6	-22.53943657	-22.53845917	-22.45823042	-22.4580127
23	F W EH Z IH G	6	-24.27911488	-24.27726633	-24.20107993	-24.20062982
24	G L EH P IH D	6	-20.36556242	-20.36579804	-20.34302202	-20.34308162
25	G L AA T IH F	6	-20.59822785	-20.59823087	-20.65641222	-20.6564051
26	G R IY B IH D	6	-20.62939192	-20.62951583	-20.67609141	-20.67611582
27	G R AA T IH F	6	-20.57694359	-20.57695004	-20.60458557	-20.60458059
28	G W IY B IH D	6	-22.83013257	-22.82897612	-22.80589101	-22.80561751
29	G W AA T IH F	6	-22.77768423	-22.77641033	-22.73438516	-22.73408229