Comparing phonotactic metrics

Connor Mayer and Austin Wagner
UCI Department of Language Science

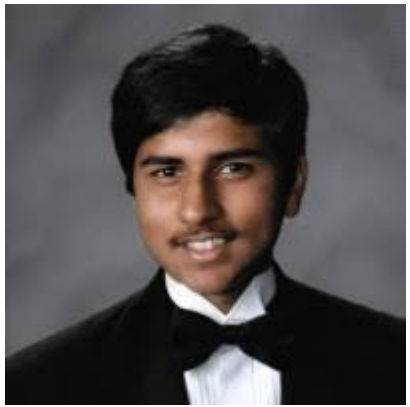
SCAMP 2025

Collaborators

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

Driven by two undergraduate research assistants

- Arya Kondur (now a PhD student in CS at UCI)
- Austin Wagner (here before you)



Roadmap

1. Background on phonotactics
2. Quantifying phonotactic probability
3. Phonotactic model bake-off, part 1
4. Phonotactic model bake-off, part 2
5. What makes a good phonotactic model?

Phonotactics

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
- /tʃknoʷ̃ɲtɕ/ is a fine Polish word, but not a good English word

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 ‘chknonch’?”
- “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

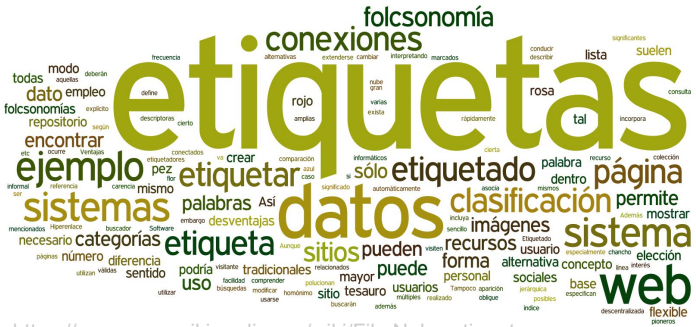
lvag

kip

What do we mean by gradience?

lvag ⋘ poik ⋘ kip

Where does phonotactic knowledge come from?



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

Generalization based on frequency

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

Lexicon

Phonotactic grammar

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 **w** composed of the segments $\mathbf{x_1 \ldots x_n}$?

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 **simple** to implement and reason about

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

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

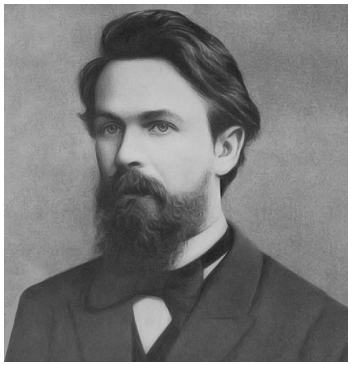
They **get us reasonably far in phonotactics**

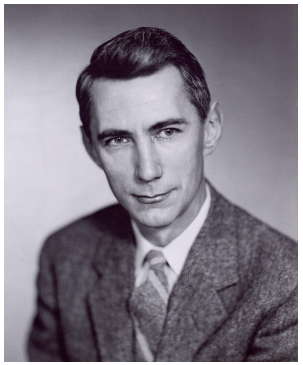
- Laplace-smoothed bigram model on English onset data ***r = 0.877***
  (Daland et al. 2011, Dai, Mayer and Futrell 2023)

Two prominent n-gram models

Researchers often use one of two n-gram models

1. **Standard n-grams** (Markov 1912, Shannon 1948)





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

Unigram and bigram models

Unigram models consider **one sound at a time**

- Do not encode information about the ordering of sounds
- A measure of *phone frequency*

Bigram models consider **two sounds at a time**

- Encode information about the ordering of adjacent pairs of sounds
- A measure of *biphone frequency*

We'll consider both of them together when evaluating models.

The standard n-gram model

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

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

Estimating probabilities from data

We can estimate probabilities by counting occurrences in a corpus (MLE)

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

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`

Smoothing

Unigrams and bigrams not observed in the corpus have probabilities of 0

Words containing unattested bigrams are assigned probabilities of 0

- This is generally undesirable

Smoothing: steal probability from attested forms and give it to unattested forms

- *Laplace smoothing*: give every unigram/bigram a default count of 1

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 w_i is the segment in the i^{th} position in word 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*

Major difference 4: (not shown) Counts in the PPC are *weighted by word frequency*

Other details about the PPC

The PPC does not use smoothing

- Not as important because probabilities are combined by addition
- Zero-probability n-grams make no contribution to the score

The PPC does not insert word boundaries

Major difference 5: The PPC does not use word boundary symbols

A priori problems with the PPC

PPC scores are not valid probabilities

Later positions run into data scarcity issues

- Lots of data to use when estimating $P(w_1 = x)$
- Much less when estimating $P(w_{10} = x)$

Can only refer to word-initial position

- w_1 is always word-initial
- The w_i corresponding to word-final position varies

Summary of model differences

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

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

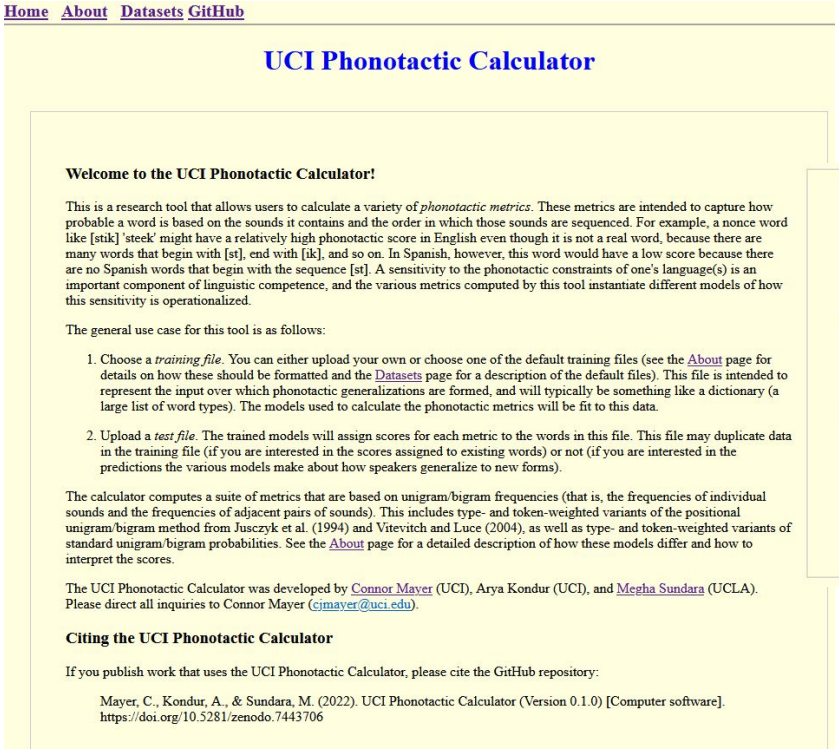
Let's compare the standard n-gram and PPC models against eight publicly available phonotactic acceptability judgment datasets

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

We also compared the effects of smoothing and token weighting on both models

- Smoothing is important for standard n-gram models, not for PPC
- Token weighting almost never improves performance for either model
- We won't present these results

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).

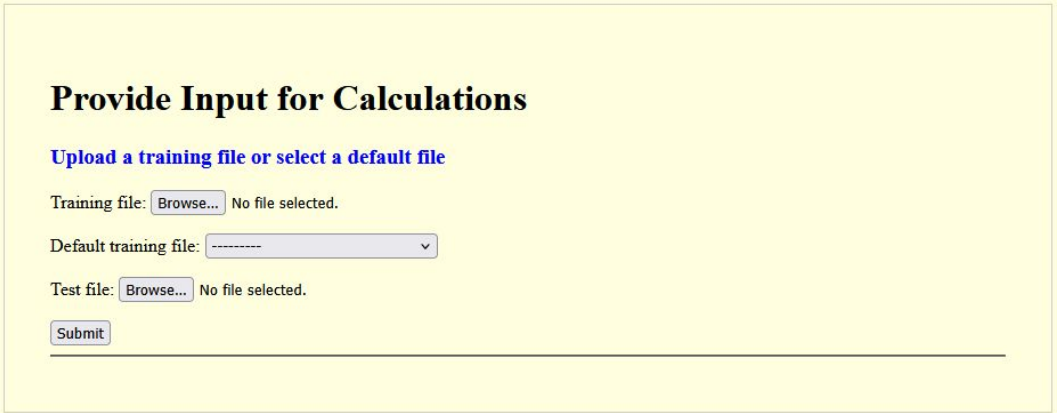
The calculator computes a suite of metrics that are based on unigram/bigram frequencies (that is, the frequencies of individual sounds and the frequencies of adjacent pairs of sounds). This includes type- and token-weighted variants of the positional unigram/bigram method from Jusczyk et al. (1994) and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

	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

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

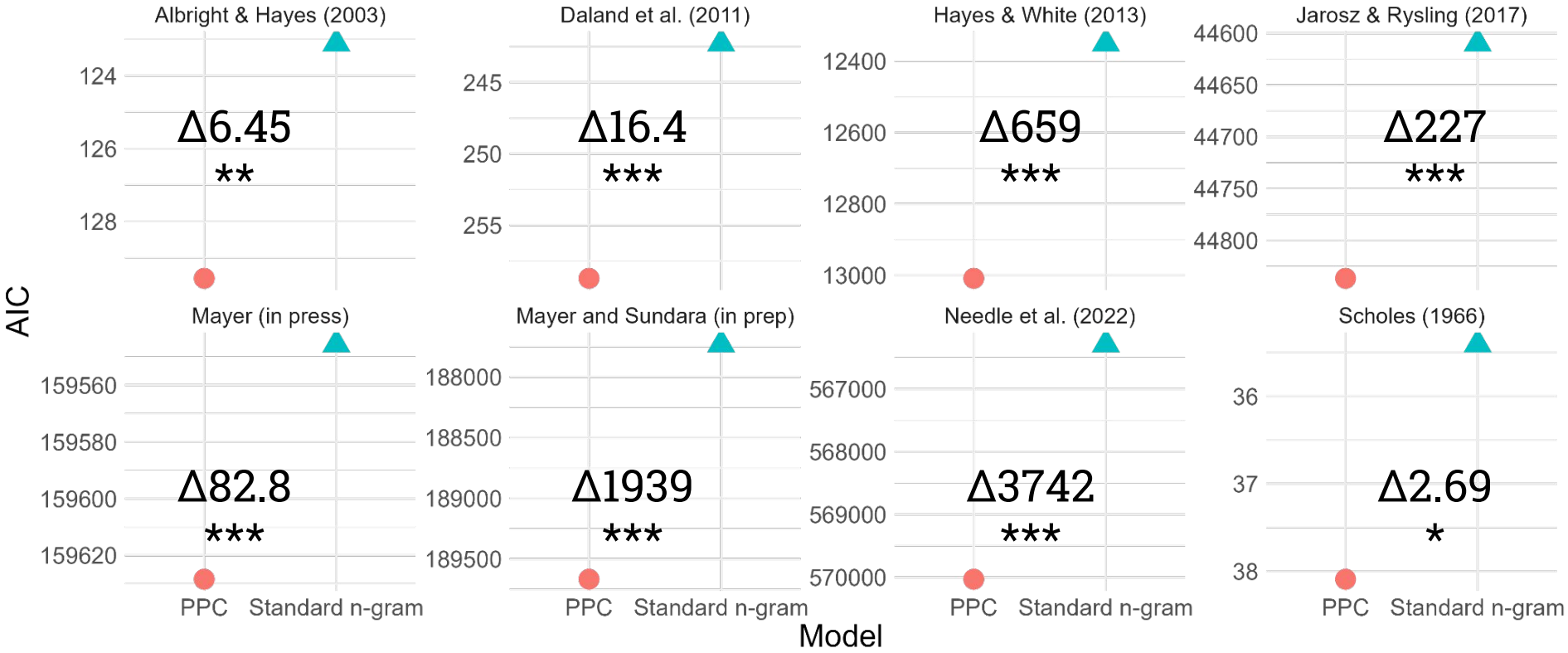
AIC is an estimate of prediction accuracy on held-out data

- We interpret AIC in terms of 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
- Increasing ΔAIC indicates increasing certainty in better model

Standard n-grams are better in every case



Model Bake-off 2: but *why?*

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?

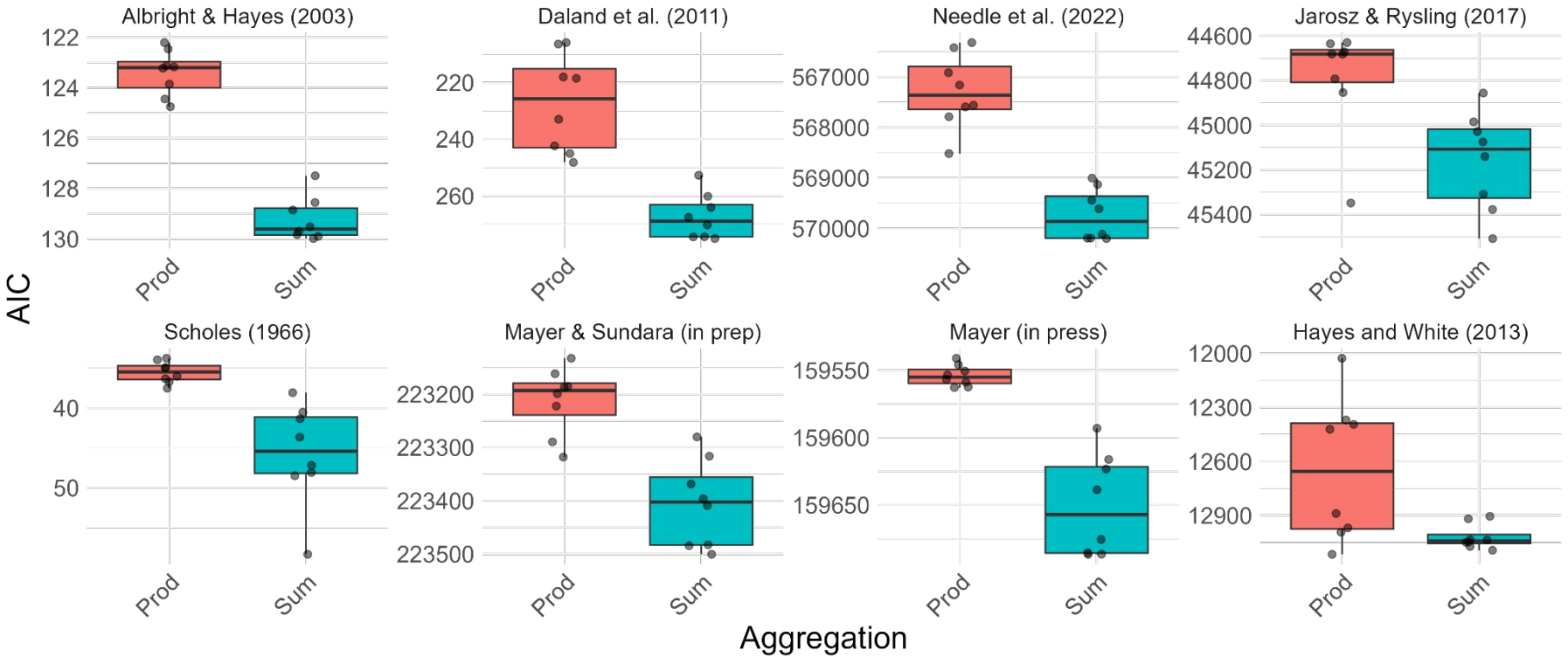
Bake-off 2 procedure

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

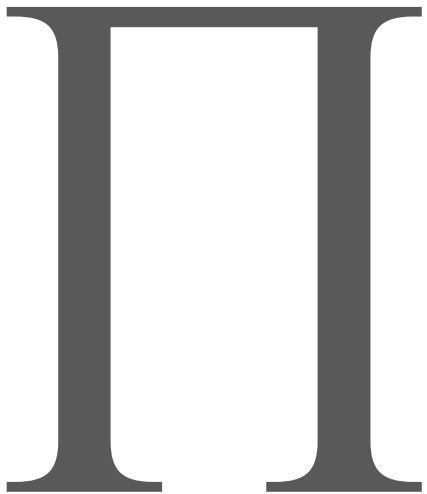
- All models were type-weighted
- Models that used product aggregation were Laplace smoothed

We fit each model to each dataset

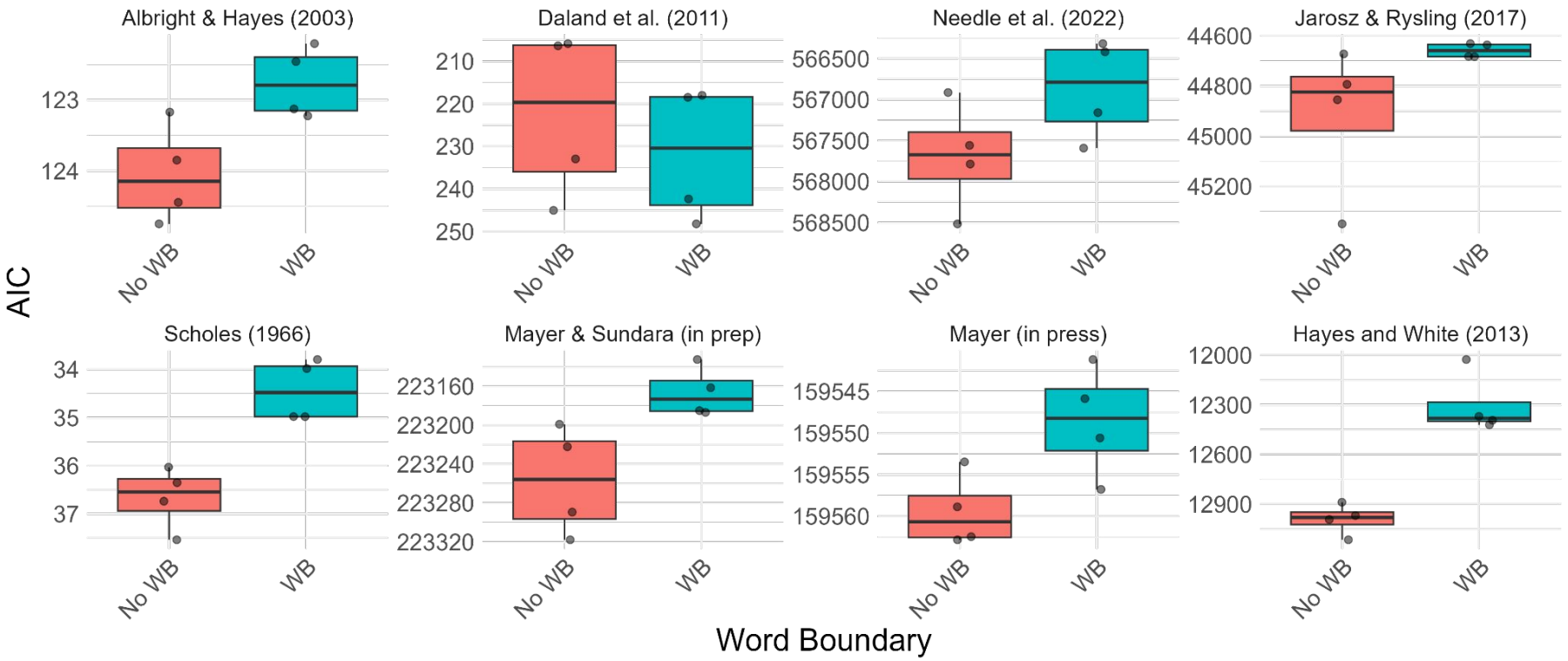
Result 1: Adding probabilities is almost always worse



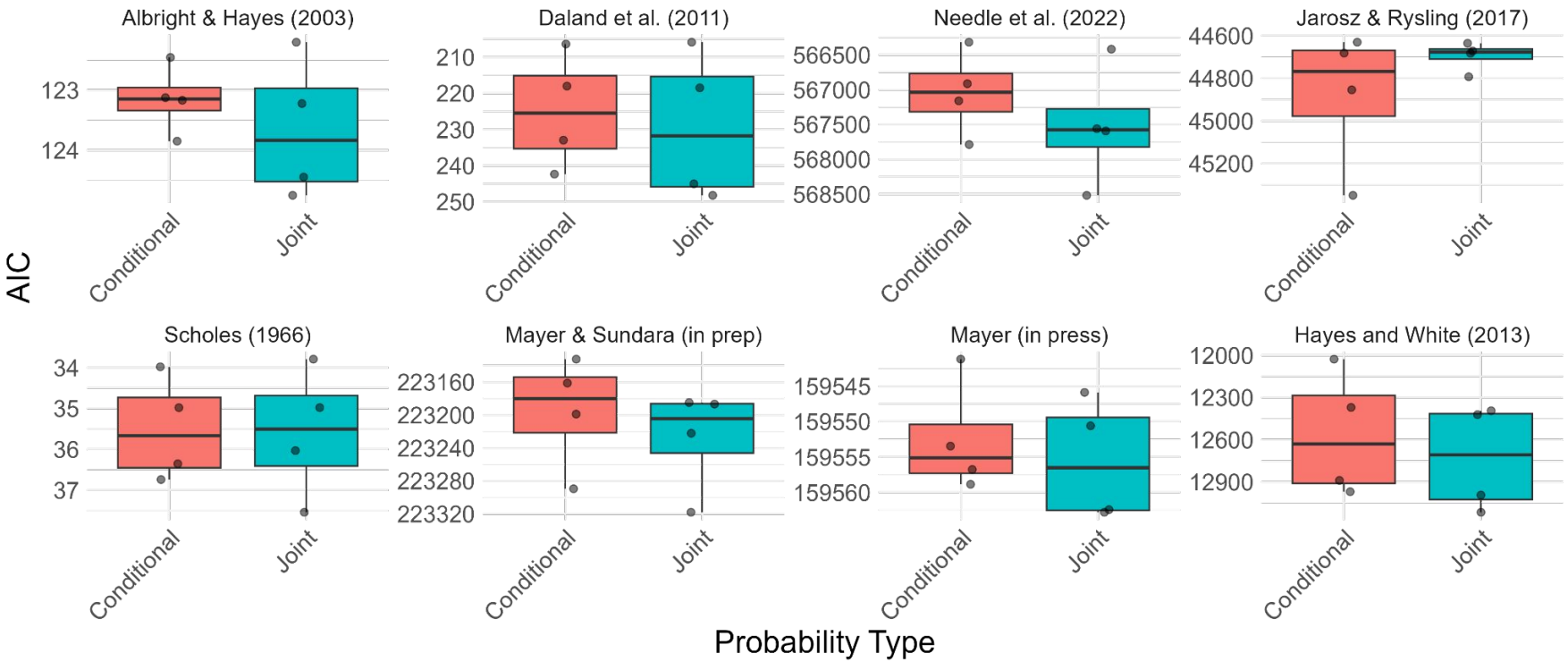
We’ll only consider the ‘product’ models going forward



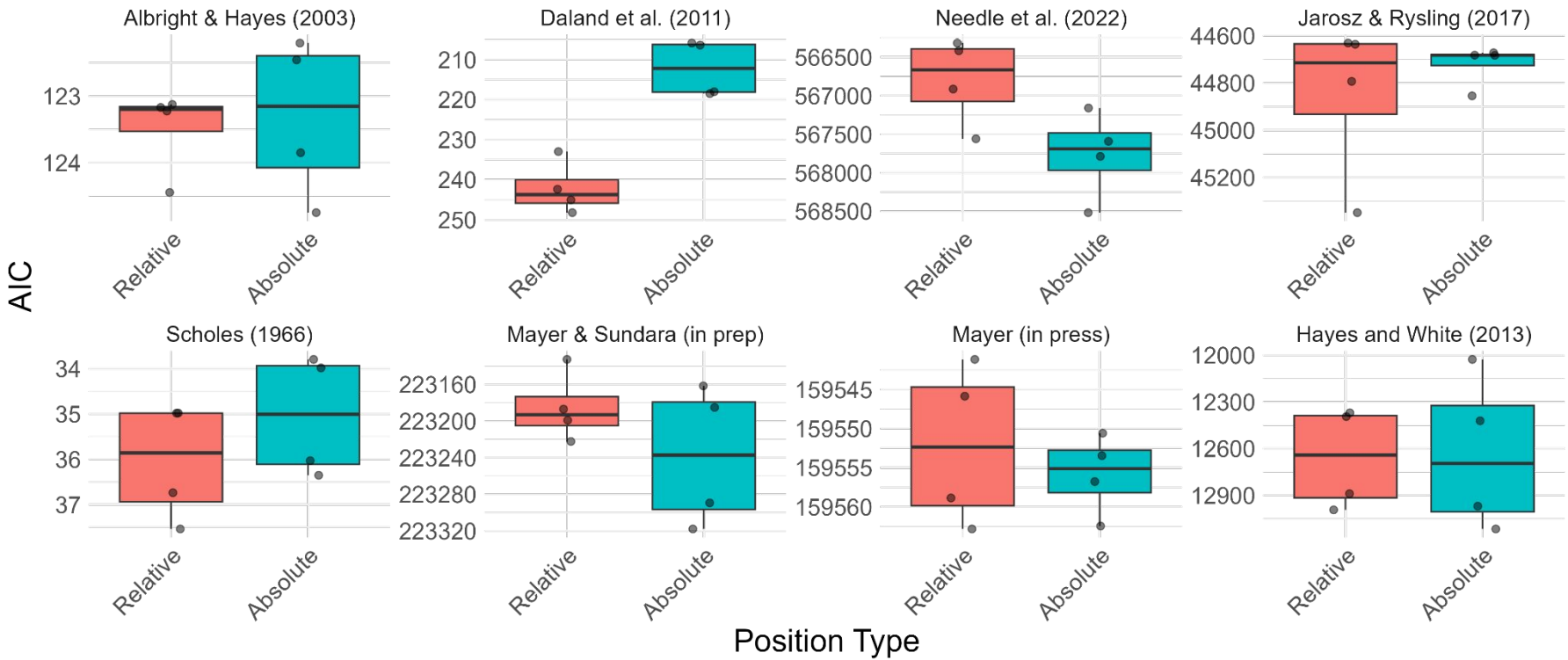
Result 2: Word boundaries help



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. **Sensitivity to 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. **Conditional probabilities > joint probabilities**
   - Difference is less substantial than aggregation or WBs
   - The two are highly correlated (Gaygen 1997, Vitevitch and Luce 1999)
   - Only conditional probabilities get us a valid probability distribution

4. **Absolute vs. relative position varied across datasets**
   - General preference for relative
   - Likely related to specific data sets used
   - E.g. bigrams can't 'see' full #CC onsets in Daland et al. (2011)
   - Positional model can track (some of) this information

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?

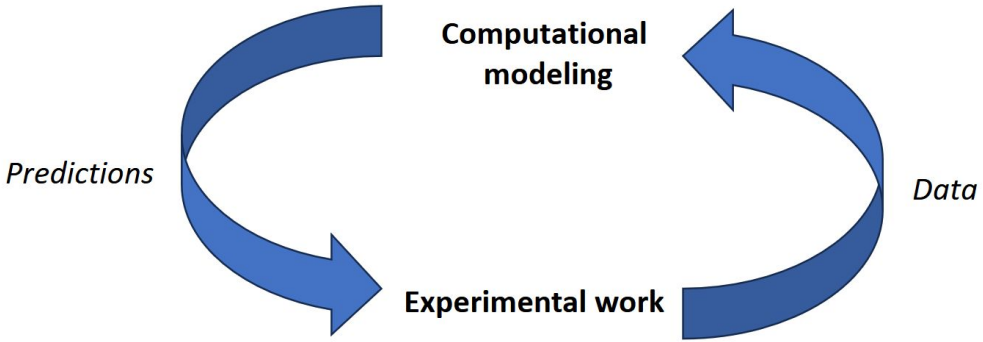
(cf. Castro and Vitevitch 2023)

Broader picture

Phonotactics are important for a range of different tasks

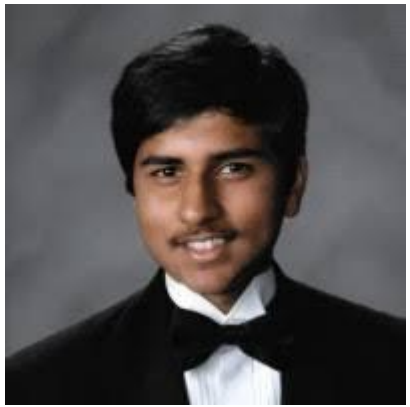
- Understanding phonotactic learning is important to understand these tasks

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



Thank you!

Megha Sundara



Arya Kondur

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.

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).

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.

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.

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 proper way to do it

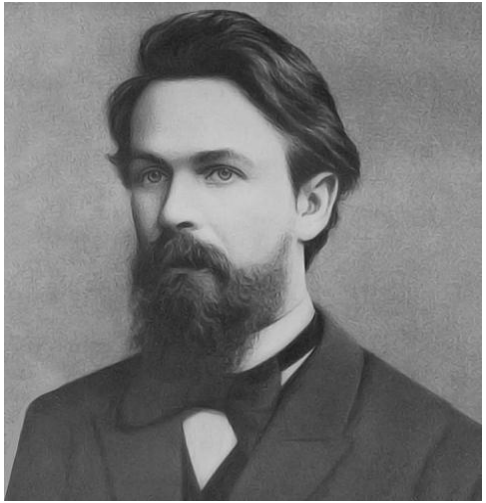
The *chain rule* of probability tells us that

$$P(w = x_1 \ldots x_n) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1x_2)\ \ldots\ P(x_n|x_1 \ldots x_{n-1})$$

This is hard to use in practice because of *data sparsity*

- Difficult to estimate probabilities from data as context gets larger

The Markov assumption



Just look at the preceding **N−1** segments

$$P(x_i \mid x_1 x_2 \ldots x_{i-1}) \approx P(x_i \mid x_{i-(N-1)} \ldots x_{i-1})$$

We'll consider *unigram* (**N=1**) and *bigram* (**N=2**) models here.

Type frequency vs. token frequency

Type frequency is consistently better

Smoothing improves performance of n-gram models but not PPC