A closer look at what/how we can learn from computational modelling of phonotactics

Karthik Durvasula

Basic Idea: Durvasula (under review) argues that the success of the MaxEnt model of phonotactics (Hayes and Wilson, 2008) in modeling word acceptability judgements is not due to gradience. He uses slight alerations of the MaxEnt model that remove gradience along different dimensions and shows that they do as good or better than the original model in predicting the acceptability of different onset clusters in English. Durvasula ultimately uses these findings to say that the success of MaxEnt cannot be used to claim that phonotactic knowledge is gradient contra the original claim of Hayes and Wilson.

A Primer on MaxEnt

A MaxEnt grammar is a phonotactic grammar and therefore only contains Markedness Constraints. The grammar simply gives the phonotactic probability of a given form in the language. Each constraint is assigned a weight, and a harmony score is assigned in the following way:

$$h(x) = \sum_{i=1}^{N} w_i C_i(x)$$

This says for N constraints, add up the weights of each specific constraint w_i multiplied by the number of violations that a word x has for that constraint $C_i(x)$.

The next step is to give it a MaxEnt value which they represent at $P^*(x) = \exp(-h(x))$. It is simply the base of the natural logarithm e raised to the negative value of the harmony score.

Suppose all possible forms are contained in the set Ω. The probability of a given form x is the maxent value of x normalized by the maxent value of all other forms in Ω.

$$P(x) = P^*(x)/Z \text{ where } Z = \sum_{y \in \Omega} P^*(y)$$

An example here may be useful. Let's assume our language can only generate the forms $\Omega = \{V, CV, VC, CVC\}$ and the MaxEnt grammar has constraints *#V with a weight of 3 and *C# with a weight of 2. The following table shows how the probabilities for each form are calculated.

x	*#V	*C#	Score	P*(x)	P(x)
V	3	0	3	0.04978707	0.04
CV	0	0	0	1	0.84
VC	3	2	5	0.006737947	0.01
CVC	0	2	2	0.1353353	0.11

These types of distributions have been supported by what Hayes et al. (2009, p. 826) refer to as **the law of frequency matching**: "speakers of languages with variable lexical patterns respond stochastically when tested on such patterns. Their responses aggregately match the lexical frequencies." Furthermore, some researchers have shown that MaxEnt style grammars are able to match corpuse frequencies (Zuraw, 2010).

Hayes and Wilson (2008) ultimate goal is to induce both the constraints and the constraint weights directly from the input that a learner receives. This is done by finding the constraints and constraint weights that maximize the probability of the data. The probability of the data is just the product of the probability of each form in the data.

$$P(D) = \prod_{x \in D} P(x)$$

The reason that this is called a Maximum Entropy grammar is because maximizing the probability of the data (maximum likelihood estimation) is the same as maximizing entropy.[1] Knowing how this occurs is not important at the moment, but if you're interested it is done by calculating a local gradient based on observed and expected violations for each possible constraint and moving in the constraint space accordingly. They also include a bias against longer constraints (both in terms of width as well as number of features used).

1 Introduction

- Speakers have intuitions about what is a possible word in their language.
- English: [bɹɪk] vs. [blɪk] vs. *[bnɪk] (Chomsky and Halle, 1968)
- In well-formedness experiments, judgements form a gradient: [brɪk] > [blɪk] > [bnɪk]
- Disagreement on the distinction between **competence** and **performance** has lead to disagreements on where to place the source of gradience in formal (computational) models.

[1] Entropy comes from Information Theory and is a measure of randomness in a system. It is given by the formula $-\sum_{x \in \Omega} P(x) \log(P(x))$.

  - Option 1: Gradience is due to performance factors and therefore outside the grammar
  - Option 2: There may be gradience due to performance factors, but the grammar itself is also gradient
  - Option 3-n: ...

- *"The key issues of interest for this paper are to understand what can be interpreted from such computational models and to understand the source of the sometimes good fit of computational models that include an underlyingly gradient phonotactic grammar (meaning, it has gradient constraints)"*

- Different types of constraints:
  - **Binary**: $w_i \rightarrow \{0, 1\}$ ... good vs. bad
  - **Scalar**: $w_i \rightarrow \{0, 1, 2, 3, \ldots\}$ ... e.g., Likert
  - **Continuous**: $w_i \rightarrow [0, 1]$ ... e.g., probabilities

- Durvasula rejects the claim by Mayer (2025) that the distinction between gradient and categorical models is a "false dichotomy". I think Durvasula's arguments here are weak and don't refute the semiring analysis provided by Mayer. This doesn't affect the paper overall, but is of interest to me.

- Strategy for comparing models:
  1. Propose a specific computational model.
  2. Train the model on a set of training data or lexicon.
  3. Simulate acceptability judgments based on the model.
  4. Test model predictions against actual speaker judgments.

- *"..to be able to infer gradience (or any other aspect of the computational model) as the source of the improved fit to human judgements, one needs an* **all things kept equal** *comparison that is missing in recent model comparisons."*

- Five main results:
  1. Fit of the MaxEnt model improves when constraint weights are binarized.
  2. Gradience in the MaxEnt model is not from constraint weights as a modified version with no constraint weights does as good or better.
  3. Gradience in the MaxEnt model doesn't stem from gradience in the training data.
  4. A degenerate MaxEnt model with only six constraints and no constraint weights does as well as the original.
  5. A minimal categorical phonotactic model is virtually identical to all other models despite being minimally specified.

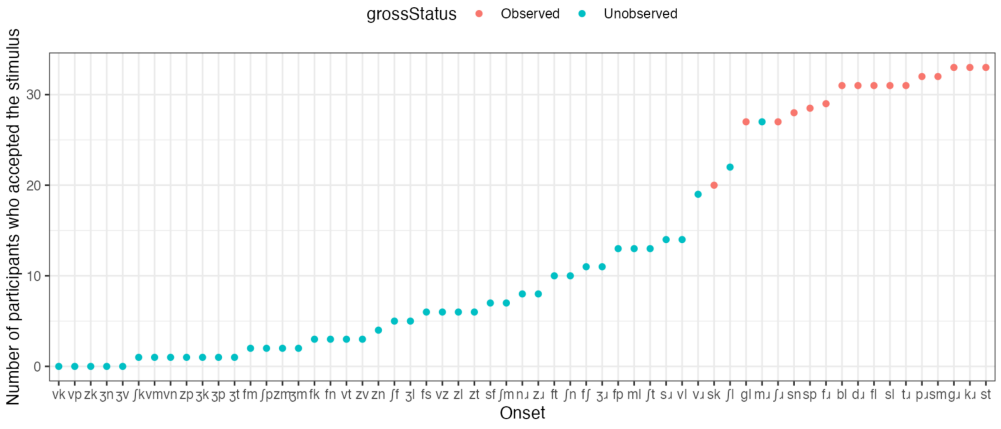
2 Modelling Decisions

- Core training data is the same as Hayes and Wilson (2008).

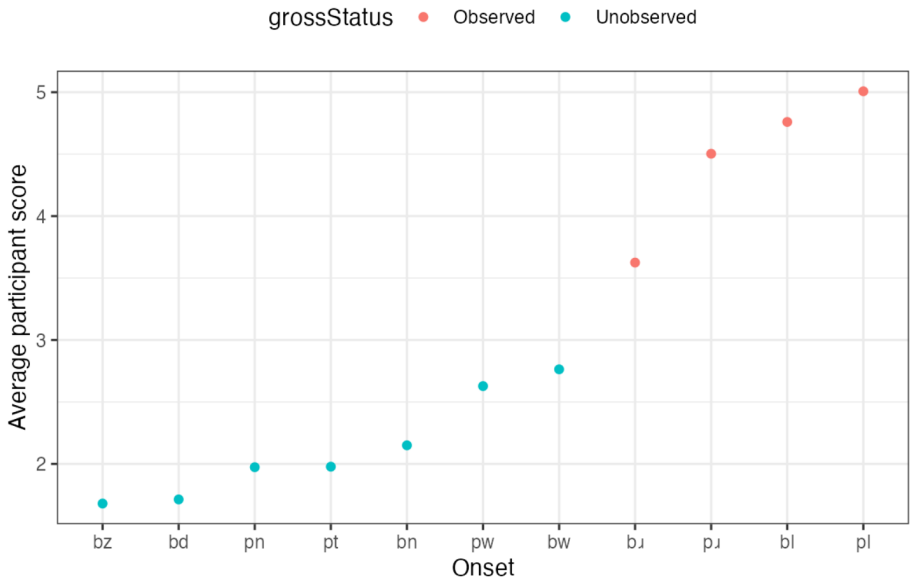
Onset	Type Freq.	Onset	Type Freq.	Onset	Type Freq.	Onset	Type Freq.
K	2,764	V	615	Y	268	Z	83
R	2,752	G	537	F R	254	S M	82
D	2,526	JH	524	P L	238	TH R	73
S	2,215	S T	521	B L	213	S K W	69
M	1,965	T R	515	S L	213	T W	55
P	1,881	K R	387	D R	211	S P R	51
B	1,544	SH	379	K W	201	SH R	40
L	1,225	G R	331	S T R	183	S P L	27
F	1,222	CH	329	TH	173	DH	19
HH	1,153	B R	319	S W	153	D W	17
T	1,146	S P	313	G L	131	G W	11
P R	1,046	F L	290	HH W	111	TH W	4
W	780	K L	285	S N	109	S K L	1
N	716	S K	278	S K R	93		

- Testing Data:

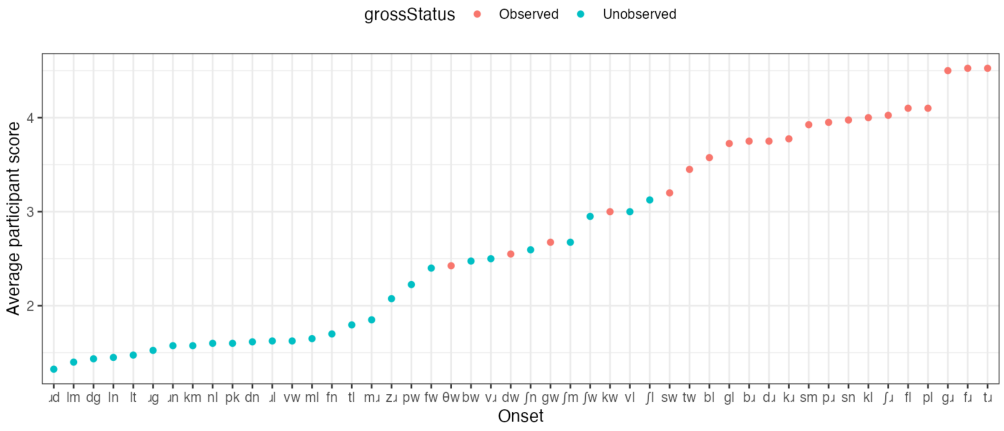
1. Scholes (1966): 35 native English speaking seventh graders who responded Yes-No to 66 monosyllabic nonwords varying in consonant sequences, where the rhymes were all bland.



2. Albright (2007): native English speakering participants who judged 30 monosyllabic nonwords varying in consonant sequences starting with p-, b-, on a 7-step Likert scale.



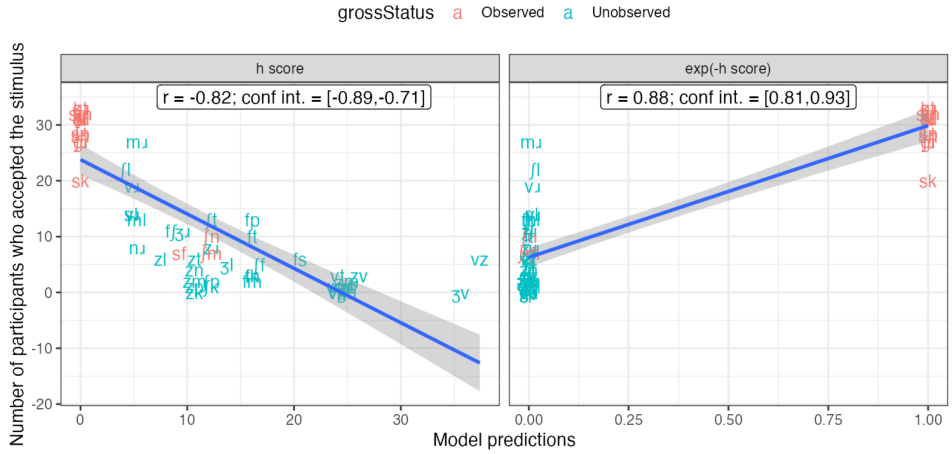
3. Daland et al. (2011): 48 native English speaking participants who judged 96 stress-initial CCVCVC non-words with 48 unique word-initial consonant sequences on a 6-step Likert scale.



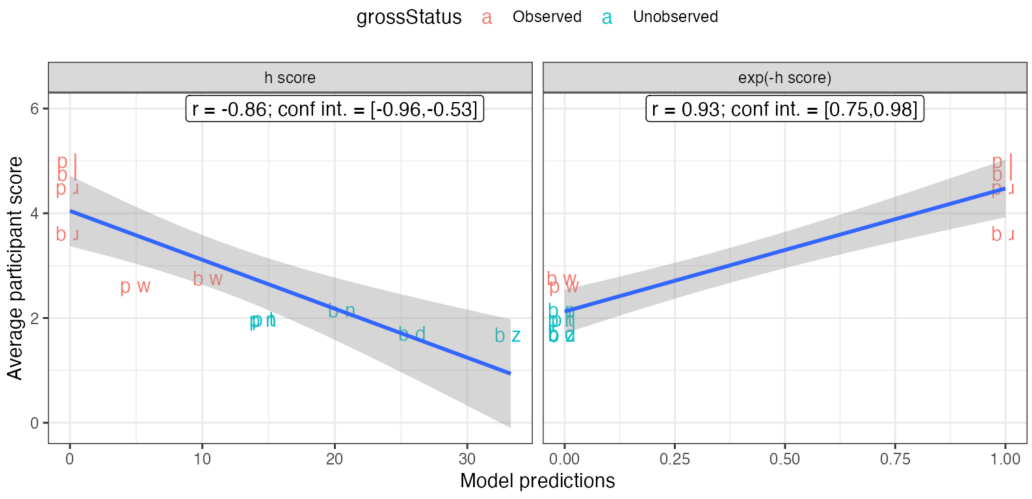
- Only onsets were judged following Hayes and Wilson (2008) and if two items shared an onset, the score for that onset was the average response across items.
- Previous studies have just shown the highest correlation value (typically using Pearson's r), but Durvasula points out that this statistic can only be used for inference if error probabilities are included. Therefore he includes the 95% confidence intervals in order to see if one fit is different from another.

3 Understanding the original Hayes & Wilson Phonotactic Model performance results

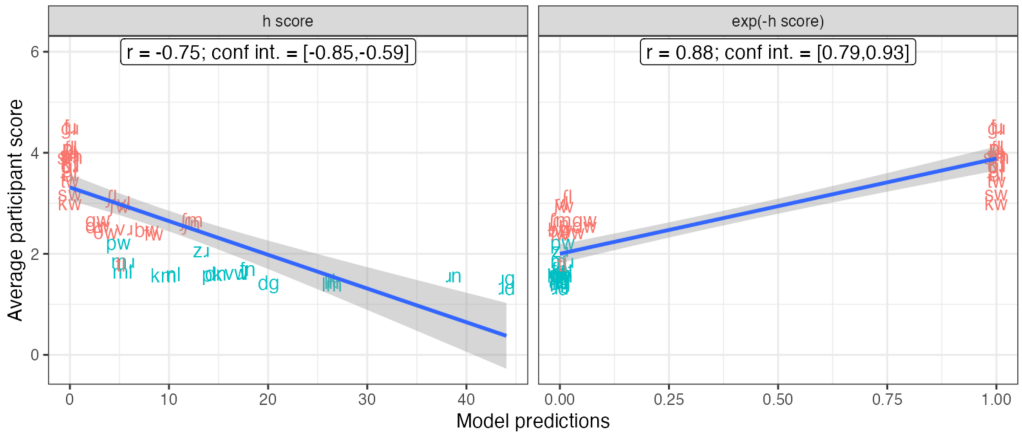
- Substantive claims in the MaxEnt model:
  1. The grammar has gradient knowledge within it (i.e., constraints are weighted)
  2. The effect of the constraints is cumulative (additive in the case of harmony score, and multiplicative in the case of the negative exponentiated scores)
  3. The predicted harmony scores are effectively binarised (a claim that will become clearer later in the paper)
  4. The probability of a particular form is in comparison to all other possible wordforms.
- **MaxEnt Model Results**
  1. Scholes (1966):



  2. Albright (2007):



3. Daland et al. (2011):

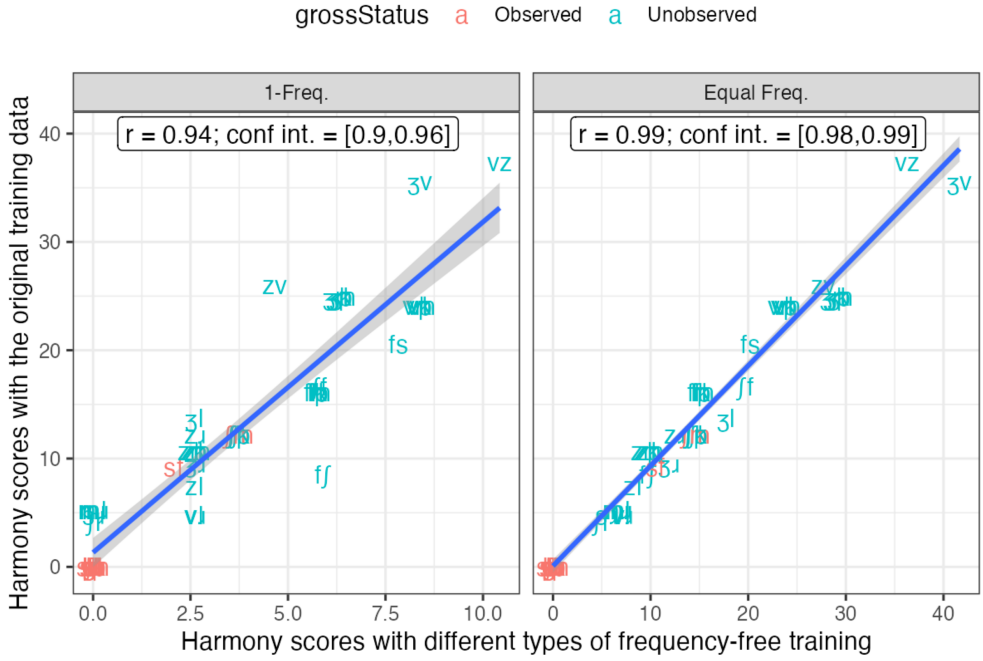


- The figures make clear that the fits are improving with negative exponentiation largely because the transform results in less gradience in the predictions.
- *"This effective binarisation of the model predictions observed in the negative exponentiated scores is our first clue that the gradiences predicted by the model is not as imporessive as what might be thought based on just looking at the model fits. It therefore raises awareness of the need for much more careful model inpsection to understand why the model is successful."*
- **MaxEnt Model without Constraint Weights Result**:
- $h(x) = \sum_{i=1}^{N} C_i(x)$
- No weight vs. original:

- The NoWeight model had better Pearsons r scores for all testing data and was statistically different in the Scholes case.

	MaxEnt	MaxEnt-NoWeights
Scholes (1966)	0.88	0.94 [0.90,0.96]
Albright (2007)	0.93	0.94 [0.79,0.99]
Daland et al. (2011)	0.86	0.90 [0.83,0.95]

- **MaxEnt Model with FrequencyFree Training Data Results**
- 1-frequency (1 of each type) vs. equal-frequency (n types / m tokens of each type)
- See table 4 (p.20) for constraints induced (this number is way lower for 1Freq training) and constraint weights (different for both FreqFree training).
- FreqFree vs. original:

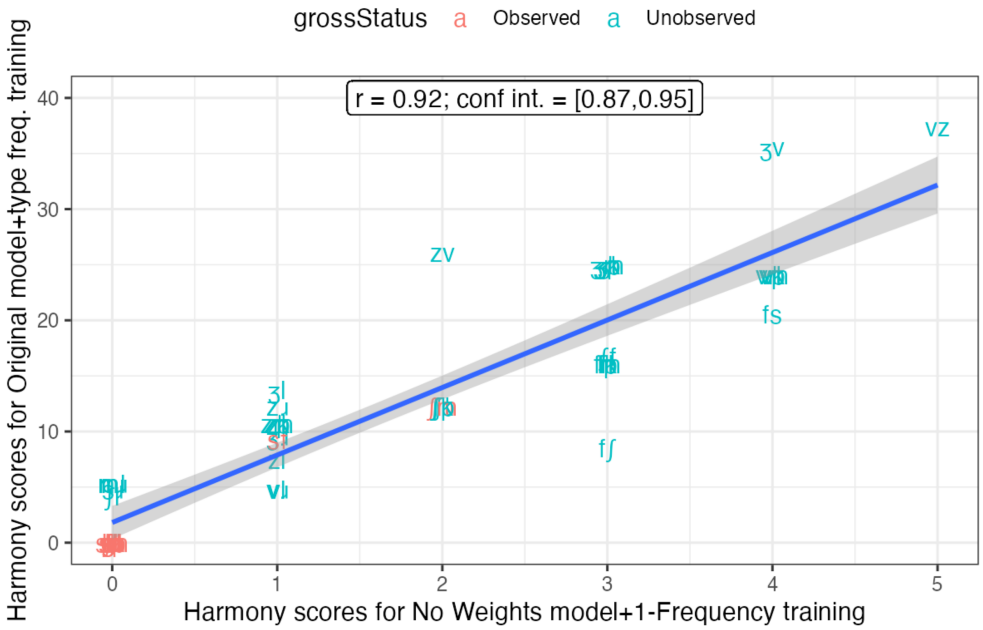


- The FreqFree models had better identical Pearsons r scores for the Schores data.

	MaxEnt	MaxEnt-EqFreq	MaxEnt-1Freq
Scholes (1966)	0.88	0.88 [0.81,0.93]	0.88 [0.81,0.93]

- **MaxEnt Model without Constraint Weights and 1Freq Training Data Results**:
- This is possibly the most extreme version of the MaxEnt model that includes no gradience in constraint weights and no gradience in training data.

- NoWeight + 1Freq vs. original:



- The original model has higher Pearson's r but is not significantly different from the NoWeight-1Freq model

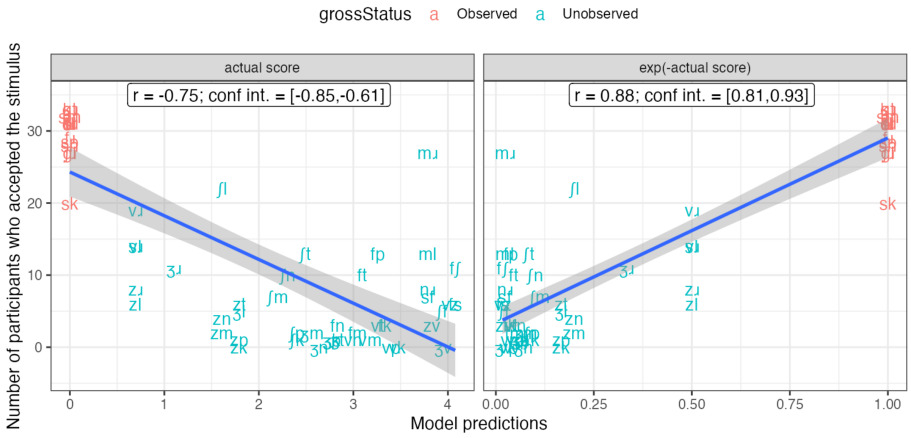
	MaxEnt	MaxEnt-NoWeight-1Freq
Scholes (1966)	0.88	0.87 [0.8,0.92]

- **Takeaways**:
  1. The presence of gradience in the predictions worsens fit
  2. The presence of gradience in the grammar (as constraint weidghts) itself worsen fit
  3. The presence of gradience in the training data doesn't affect fit

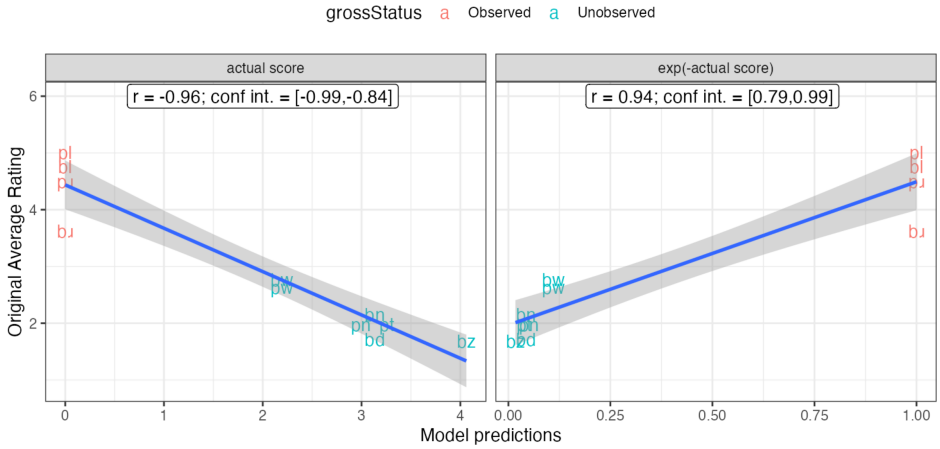
4 The minimal (categorical) phonotactic model

- The goal here is to show gradience in acceptabilities without any gradient generalizations.
- The model: All the observed single feature unigrams/bigrams and segment unigram/bigrams in the learning corpus form a **positive grammar**
- A theory of *performance*: For each nonsense word tested, one counts the number of featural and segmental unigrams/bigrams that are not in the grammar and takes the log of this value (with add one corection).

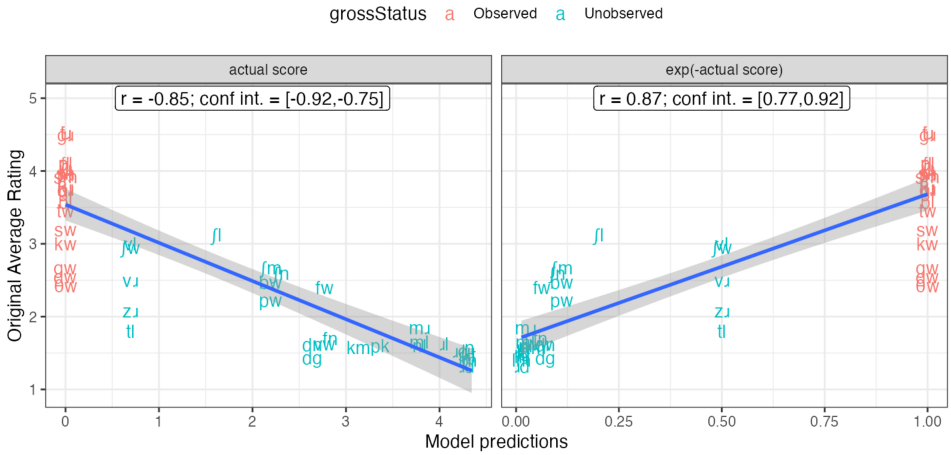
1. Scholes (1966):



2. Albright (2007):



3. Daland et al. (2011):



- The MC Model was not statistically different from the MaxEnt model for harmony

or exp scores

	MaxEnt	MC-Harmony	MC-Exp
Scholes (1966)	-0.75/0.88	-0.75[-0.85,-0.61]	0.88 [0.81,0.93]
Albright (2007)	-0.83/0.94	-0.96 [-0.99,-0.84]	0.94 [0.79,0.99]
Daland et al. (2011)	-0.73/0.86	-0.85 [-0.92,-0.73]	0.87 [0.77,0.92]

- Other than Scholes, the exponentiation doesn't really change the performance of the MC Model.
- Despite being minimal, this model does at least as well as the MaxEnt model.

5 Conclusion

- It isn't phrased this way in the paper, but the argument that Durvasula is making relates to inference procedures when evaluating experimental/behavioral results in relations to the underlying system. I write about this in my dissertation (Nelson, 2024), but I was really influenced by Guest and Martin (2023). In relation to cognitive science more broadly they critique LLM research that "affirms the consequent". This comes from mathematical logic when you have the proposition $T \rightarrow D$. Suppose T stands for the claim that our theory does what humans do and D stands for our theory being correlated with the data. What's of interest here is that D cannot prove T to be true in this case. Instead, all that can happen is $\neg D$ (our theory not being correlated with the data) can prove T to be false.
- This is all made worse by *auxiliary assumptions* which Durvasula references multiple times throughout the paper. This idea goes back to Duhem (1954) and Quine (1951) and basically tells us that if the data doesn't correlate with our theory, it could be the (core) theory that's wrong or it could be some auxiliary assumption such as a linking hypothesis between data and theory. Lakatos (1970) is a beautiful paper that discusses this in great detail.
- *"The bottomline is that if we are interested in learning about the undelrying competence/-grammatical system, we need to be as critical as possible, and only allow new/more constructs into the system if they are theoreticaly unavoidable and can be justified through repeated experimentation an analysis – the simple fact of just accounting fo the data is far from sufficient to make the case."*
- Theorists and computational researchers have to performance facst seriously and incorporate them into their models/experiments.
  - This is an old view. See for example what Mohanan (1986, p. 183; emphasis original) wrote 40 years ago:
    "Practitioners of phonology often distinguish between *internal* evidence, which

> consists of data from distribution and alternation, and *external* evidence, which consists of data from language production, language comprehension, language acquisition, psycholinguistic experiments of various kinds, sound patterning in versification, language games, etc. [...] The terms "internal" and "external" evidence indicate a bias under which most phonological research is being pursued, namely, the belief that the behaviour of speakers in making acceptability judgments is somehow a more direct reflection of their linguistic knowledge than their behaviour in producing language, understanding language, etc. This bias appears to be related to the fact that linguistic knowledge is only *one* of the inputs to language production, language comprehension, and other forms of language performance. What accounts for the facts of performance is a *conjunct* of a theory of linguistic knowledge ("What is the nature of the representation of linguistic knowledge?") and a theory of language performance ("How is this knowledge put to use?")."

References

Albright, A. (2007). Natural classes are not enough: biased generalization in novel onset clusters. Ms., Massuchesetts Institute of Technology.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. ERIC.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., and Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, 28(2):197–234.

Duhem, P. M. M. (1954). *The Aim and Structure of Physical Theory*. Princeton University Press.

Durvasula, K. (under review). A closer look at what/how we can learn from computational modelling of phonotactics. Unpublished manuscript.

Guest, O. and Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, pages 1–15.

Hayes, B., Siptár, P., Zuraw, K., and Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language*, pages 822–863.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Lakatos, I. (1970). *Falsification and the Methodology of Scientific Research Programmes*, volume 4, page 91–196. Cambridge University Press.

Mayer, C. (2025). Reconciling categorical and gradient models of phonotactics. *Proceedings of the Society for Computation in Linguistics*, 8(1).

Mohanan, K. P. (1986). *The theory of lexical phonology*, volume 6. Springer.

Nelson, S. (2024). *The Computational Structure of Phonological and Phonetic Knowledge*. PhD thesis, Stony Brook University.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, 60(1):20–43.

Scholes, R. J. (1966). *Phonotactic grammaticality*. de Gruyter.

Zuraw, K. (2010). A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language & Linguistic Theory*, 28:417–472.