Testing Models in English Phonotactics

Jiayi (Gemma) Huang, BA Honors Linguistics; Prof. Timothy O'Donnell, Depart. of Linguistics
McGill University, Montreal Computational & Quantitative Linguistics Lab

Abstract

The adoption of words is governed by systematic intuitions on the likelihood of different combinations of sounds in a language. For example, between two hypothetical English words "blick" and "bnick", "blick" has a higher likelihood to be accepted as an English word by a native speaker (Chomsky and Halle, 1965). Understanding such constraints on allowable sound sequences is crucial to understanding language productivity and second language acquisition. The focal point of the project to explore and test models of phonotactics from three major theoretical traditions. Three classes of models of phonotactics are being compared: maximum entropy models, Bayesian models, and neural-network-based models.

Methods and Objectives

This experiment is a simple judgment task. The setup of the experiment is as follows: the experiment prompts the experiment-taker with a series of made-up English words mixed with real English words along with their pronunciation (audio stimuli). We ask the participants to rate how acceptable they think the made-up word is on a scale of 1 to 5: 5 being the most acceptable, and 1 being the least. This experiment is implemented using JavaScript, specifically the jsPsych package. There are 17 million raw stimuli, and they are in the form of list of ARPAbet strings and IPA that represents the pronunciation of the pseudo words. The list of IPA is converted to the corresponding audio files using Amazon Polly. This experiment will be hosted on Amazon Mechanical Turk using psiTurk.

Figure 1. An example of the experiment interface

The main objective of the experiment is to collect data that will allow us to compare multiple automatic phonotactic scorers. A good phonotactic scorer should reflect how humans judge the acceptability of a fake word. Thus, the results of the experiment will expose the weaknesses and strengths of various automatic scorers.

Background

Classical Model:

- The key framework: three levels of adequacy of a grammar.
- **The *observational adequacy*** (the first level): a grammar that describes the data on which it is based, and nothing more.
- **The *descriptive adequacy*** (the second level): a grammar that gives a correct account of the speaker's "tacit knowledge", which also describes the acceptable pseudo words in a language.
- **The *explanatory adequacy*** (the third level): a grammar that provides a principle basis for the selection of descriptively adequate grammars

Maximum Entropy Model:

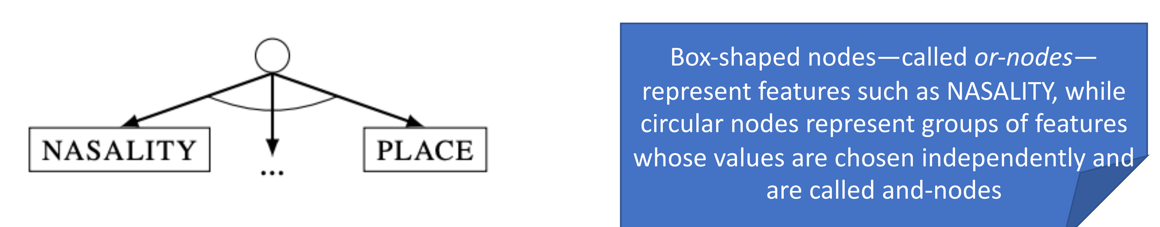
- A grammar is assigned numerical weights in terms of its constraints.
- A higher weight lowers the probability of the form that violates the constraints.
- The score of each word is the sum of the products of constraint violations and constraint weight:

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

- The model we will be testing is *BLICK (*Bruce Hayes and Colin Wilson, 2012*)*

Bayesian Model:

- Uses Bayes' theorem to update and compute probabilities after obtaining new data via prior, posterior and likelihood functions.
- In this project, we will be testing the generative model proposed by Richard Futrell (Futrell *et al, 2017)*.
- The generative model views the problem of learning the phoneme inventory as one of concentrating probability mass on those segments which have been observed before.
- This model combines the AND-OR graph and the technique of stochastic memorization (remembering probabilistic events that have already occurred).

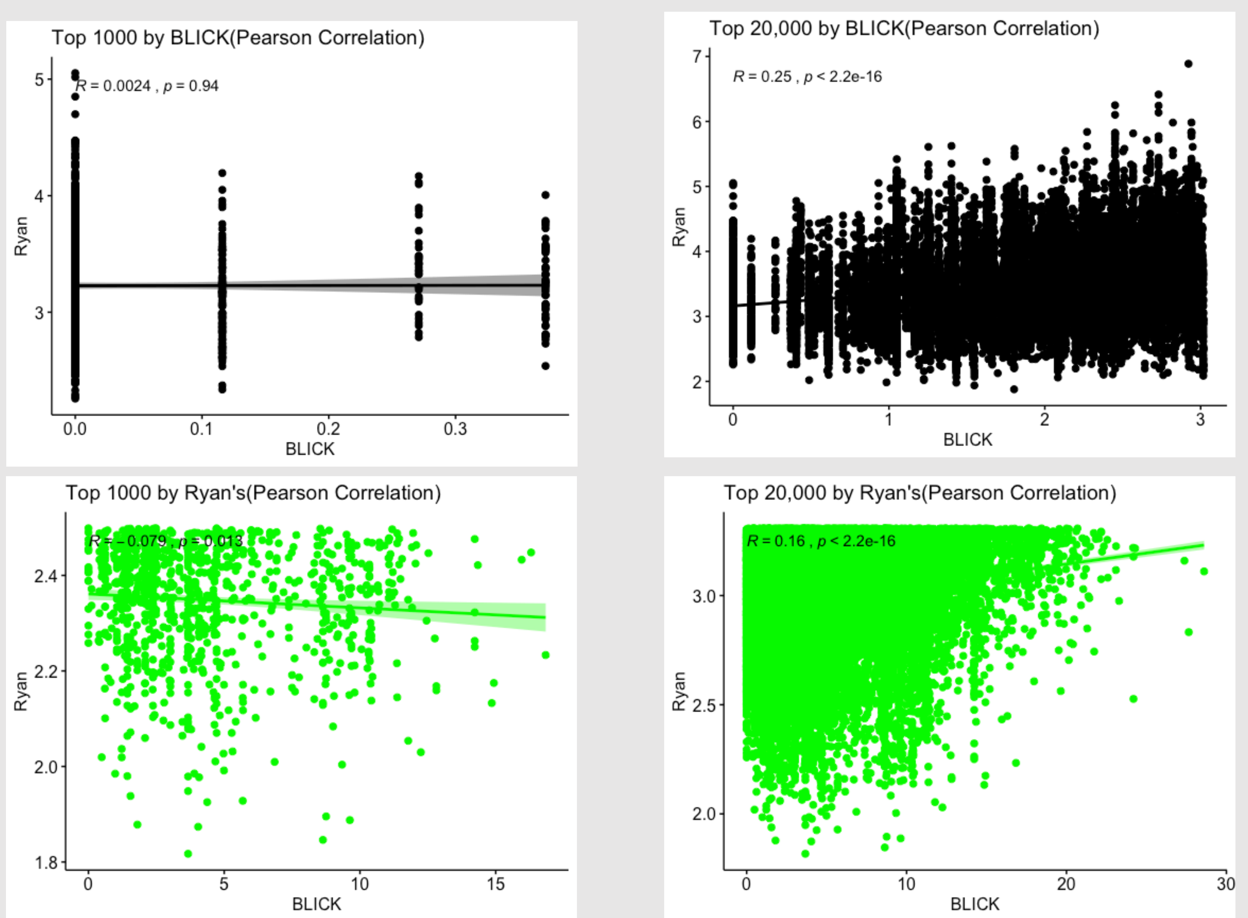


Neural Model:

- Neural sequence to sequence models have been used to capture phonotactic generalizations
- The distinction between core features and combination features: A core feature refers to a single indicator, such as whether a certain phoneme is present; a combination feature describes a conjunction of features occurring together.
- The neural model we will be testing is provided by Shijie Wu and Ryan Cotterell.

What did we find?

- Due to time constraints, we were only able to obtain scores from the BLICK model and Ryan's neural network model.
- The Pearson correlation coefficients of the two scores for the words that scored top 1000 and 20,000 by BLICK and Ryan's model are shown in the graphs below.
- The results suggest that the correlation between the two model is weak, which posts a challenge for selecting the test stimuli for the experiment.
- At the moment, we have chosen the fake words that scored top 1000 by Ryan's model as the stimuli for the prototype.



Going forward

- Though we did not have enough time to collect human judgements, the implementation of the experiment and a pilot of the project is ready to be deployed.
- The next steps will be:
  - Run Richard's Bayesian model for phonotactics on our stimuli corpus and get the scores.
  - Given the scores from all three models, perform statistical analysis to help us better decide what words are most likely to be discriminated by all three models.
  - Once the human judgements are collected, further statistical analyses will be required.

Acknowledgments

Thank you to Professor O'Donnell for this opportunity and for his guidance throughout this internship. I am also grateful for all the help from my lab members Vanna Willerton, Yves Blain-Montesano and Seara Chan. I greatly appreciate the assistance and funding provided to me by the Arts Internship Office.

Contact

[name] Jiayi Huang
[organization] McGill University
[email]: jiayi.huang@mail.mcgill.ca

References

1. Adam Albright. 2009. *Feature-based generalization as a source of gradient acceptability.*
2. Bruce Hayes and Colin Wilson. 2008. *A maximum entropy model of phonotactics and phonotactic learning.*
3. Bruce Hayes and Colin Wilson. 2012. *Blick - a phonotactic probability calculator*
4. Futrell, R., Albright, A., Graff, P., and O'Donnell, T. J. (2017). *A generative model of phonotactics.*
5. Noam Chomsky and Morris Halle. 1965. *Some controversial questions in phonological theory.*
6. Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing Ch. 3, 4, 5 ,9*