Evaluating Wasserstein GAN Discriminators as Models of Well-Formedness judgments[*]

Bruno Ferenc Šegedin
Brown University

1 Introduction

A central goal of phonological theory is to explain how speakers arrive at judgments of well-formedness (e.g. Hayes & Wilson, 2008; Breiss & Albright, 2022; Pierrehumbert, 2001b). These judgments are often modeled as mappings from linguistic forms to values like counts or judgment ratings that reflect the extent to which certain patterns are preferred or dispreferred; such models can take the form of weighted constraint models, which formalize well-formedness judgments as weighted sums of constraints (e.g. Hayes & Wilson, 2008).

While much theoretical and computational work has focused on modeling well-formedness judgments over discrete symbolic representations, human learners ultimately acquire phonological generalizations from continuous acoustic input. This paper asks whether a neural network trained only on raw speech, and with no explicit phonological supervision, can nevertheless produce outputs that resemble well-formedness judgments. As a starting point, we investigate the discriminator in a WaveGAN model. The discriminator's architecture maps waveforms to scalar values (referred to here as "discriminator scores") and this study explores whether, as in weighted constraint models, these scores can adhere to meaningful structures in the input. If successful, such a model would offer a proof of concept that the necessary building blocks for learning phonological well-formedness are available directly in acoustic experience.

2 Background

2.1 *Linear models in phonology* Many influential frameworks formalize phonological knowledge as some function that assigns a scalar score to an input form. In Optimality Theory and its variants, grammaticality is determined by weighted or ranked constraints, while in Harmonic Grammar and Maximum Entropy models, well-formedness corresponds to a ranking or weighted sum of constraint violations (e.g. Prince, 2004; Mayer & Nelson, 2020; Mayer, 2025). In such models, the grammar is "linear" in the sense that each constraint contributes additively to an overall score, and the relative strength of these constraints is captured through weights that scale the relative contribution of each constraint to the output scores.

Such models have shown success in accounting for gradient well-formedness judgments (e.g. Breiss & Albright, 2022) and probabilistic patterns in phonotactic count data (e.g. **?**). However, they rely on hand-designed constraints or features defined over discrete representations. This leaves open the question of how learners build up representations and constraints from their raw linguistic input to begin with.

2.2 *Deep neural networks in phonology* Recent work has demonstrated that deep neural networks trained on speech can capture a range of phonetic and phonological regularities, including phoneme categorization, phonotactic patterns, and phonological alternations (e.g. Beguš, 2020a, 2022; Baevski et al., 2020; Barman et al., 2024; Shain & Elsner, 2019, 2020; de Heer Kloots & Zuidema, 2024; Taigman et al.,

[*] I thank Gašper Beguš and members of the Berkeley Speech & Computation lab for feedback and discussions related to this project.

Proceedings of *AMP 2025*

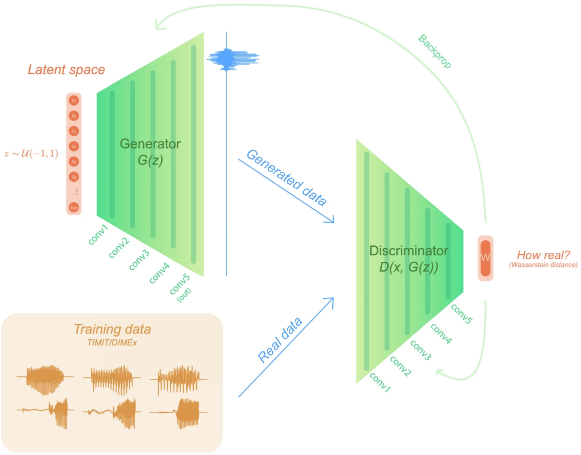

Figure 1: WaveGAN architecture used in this study taken from WaveGAN architecture used in this study, taken from Beguš et al. (2023)

2016). Unlike linear constraint-based models where constraints are pre-defined, theoretically motivated, and characterize discrete strings of symbols, neural networks can operate directly over high-dimensional acoustic inputs and learn intermediate representations that abstract away from surface variability. The ability of models to capture relative phonotactic well-formedness from acoustic data has only recently begun to be explored (e.g. de Heer Kloots & Zuidema, 2024; Gauthier et al., 2025). de Heer Kloots & Zuidema (2024), for example, test whether Wav2vec2 can resolve phonetic ambiguity in the direction of a phonotactically more well-formed utterance during speech-to-text transcription. They find that a phonetically ambiguous /r/-/l/ segment is classified as /r/ when it follows a /t/ and as /l/ when it follows an /s/, showing that models can leverage phonotactic knowledge. However, such evidence depends on fine-grained comparisons of minimal-pair-like acoustic patterns, and this approach cannot be readily deployed to produce values corresponding to the well-formedness of any input. Further, many neural models of phonology (including de Heer Kloots & Zuidema (2024)) are either supervised classifiers or in the case of self-supervised models, require probing to evaluate whether models cluster acoustic inputs in phonologically substantive ways (e.g. Gauthier et al., 2025). The question remains whether a neural network can produce interpretable, scalar values as judgments of phonological well-formedness, and whether it can learn this solely from exposure to unlabeled acoustic data remains to be determined.

3 Current approach: GAN discriminators as sources of well-formedness judgments

The goal of the current study is to test whether a WaveGAN discriminator can serve as a phonotactic learner that can assign well-formedness judgments to any input, and which requires no ground linguistic labels. WaveGAN is an example of a generative adversarial network or "GAN" Goodfellow et al. (2014). GANs typically consist of a generator that is trained to map low-dimensional random input values to high dimensional representations like audio waveforms or images. The GAN training paradigm can be informally summarized as follows: The discriminator's task is to correctly classify real training examples and waveforms produced by the generator. The generator maps a compact set of random input values to a waveform output, and its goal is to "fool" the discriminator and achieve as high a score as possible on its outputs. While a recent and extensive body of work has focused on how the generator encodes linguistic information in its latent space (e.g. Beguš, 2020b, 2021), the discriminator is typically regarded as an auxiliary model and its learned linguistic representations have received relatively little attention.

This study explores whether a discriminator trained on raw acoustic evidence for vowel identity can systematically differentiate vowel identity from vowel disharmony patterns, despite receiving no explicit information about phonological categories or constraints. From the standpoint of phonotactic modeling, the

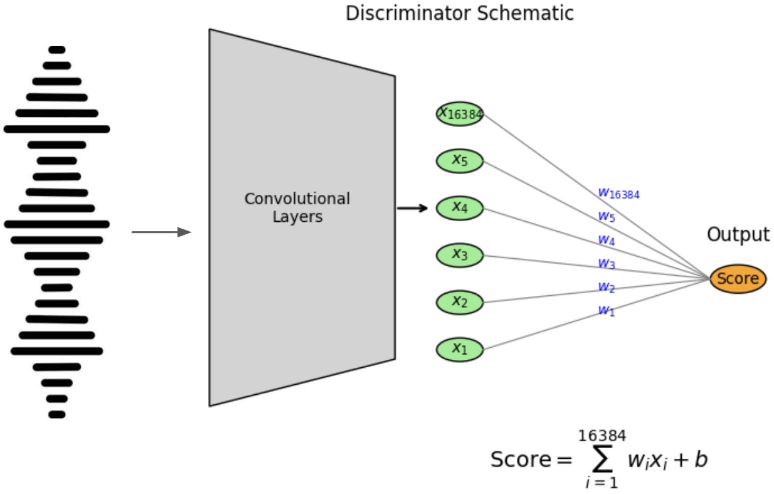

Figure 2: A simplified schematic of the discriminator. This is meant to illustrate that the convolutional layers map an acoustic output to a list of feature values- 16,384 values in the original WaveGAN model. The output discriminator score is then a weighted sum of these feature values, where the weights are parameters learned from data. As in any classification task feature could in principle be thought of as reflecting some generalization about the input, and the weights can reflect learned "preferences" or "dispreferences" for those input characteristics.

discriminator is at least superficially appropriate because its architecture forces it to map any high dimensional time-series or acoustic data, to a single continuous unbounded value, the "discriminator score." Under the Wasserstein GAN training objective, the discriminator (often referred to as a "critic" in Wasserstein training), is optimized to maximize the "Wasserstein distance"- the difference in *mean* scores it assigns a batch of training examples and a batch generated examples. Critically, even training examples or training-like stimuli are likely to follow a large distribution of scores, meaning that the "critic" learns to reward some aspects of the stimuli more than others as opposed to outright classifying real and generated examples. The ultimate goal of this study is to characterize the variability in the discriminator's scores and assess whether it reflects linguistically meaningful differences or abstractions among stimuli.

The discriminator is thus appealing as a candidate unsupervised model of phonological well-formedness judgments for three reasons: (1) it defines a mapping from an input signal to a single scalar value. (2) this mapping is learned without ground-truth labels, relying only on exposure to the distributional structure of the training data, and (3) the final layer of the discriminator computes a weighted sum of learned features, formally resembling linear constraint-based models of grammaticality, and which could in principle encode constraint-like generalizations about the input.

4 Experiment

The goal of the experiment is to test the extent to which a WaveGAN discriminator trained on raw speech assigns systematically different scores to harmony (identity) versus disharmony patterns, despite receiving no explicit information about phonological categories or constraints during training.

4.1 *Data & training* This study has two training conditions: identity and disharmony. The data in each condition consist of 2,000 Amazon Polly CVCVC words, with vowels sampled from the set [/i/, /a/], and the stops consisting of voiceless fricatives, stops and affricates. The identity condition consists of 2,000 words where either both vowels are /i/ or both vowels are /a/. The disharmony condition consists of 2,000 words where identity is prohibited, thus where /i/ and /a/ must co-occur, in either order. The 2,000 words in each condition can be broken down into 500 unique CVCVC sequences all generated by unique Spanish voices generated by Amazon Polly's neural text-to-speech engine. There are two male voices (Sergio and Pedro) and two female voices (Lucia and Lupe).

4.2 *Architectural Manipulation* The standard WaveGAN discriminator computes its final output as a weighted sum over 16,384 feature values (1024 channels by 16 time steps). While this weighted summation formally mirrors linear constraint-based models, the number of "constraints" is implausibly large from a cognitive perspective.

This study thus also explores whether reducing the number of "constraints" yields interpretable and constraint-like patterns in the activation. We train variants of the discriminator with progressively fewer output channels, resulting in fully connected layers of size 1024×16 (baseline), 32×16, 8×16, and 2×16. All other architectural and training parameters are held constant. A possible prediction is that with fewer output features, individual output features may align more closely with abstract properties such as identity versus disharmony, yielding disentangled "constraint-like" output features that clearly discriminate identity and disharmony, even if such discrimination does not emerge just based on the discriminator score.

4.3 *Evaluation* Each discriminator is tested on its ability to assign higher scores to harmony patterns given training on identity, and separately to assign higher scores to disharmony patterns given training on disharmony. In each condition, the discriminator is evaluated on held out data. We evaluate the correspondence between discriminator scores and attributes of the input data by fitting the following mixed-effects model on distributions of scores separately for each of the 8 discriminators (2 training conditions x 4 bottleneck sizes):

$$\texttt{Score} \sim \texttt{Identity}+\texttt{Identity}:\texttt{V1}+\texttt{V2}+(1+\texttt{Identity}+\texttt{V2}+\texttt{Identity}:\texttt{V1} \,\|\, \texttt{Speaker}) \quad (1)$$

These regressions are implemented using the lmertest package (Kuznetsova et al., 2015) in R.

5 Results

5.1 *Discriminator Score Analysis* Each model is trained for 1000 epochs (31,000 training steps), at Brown University's Center for Visualization and Computation. Table 1 shows the estimates for the identity coefficient in both training conditions across all 4 bottleneck sizes. Smaller bottlenecks do not facilitate learning of the pattern: the classification of the model with a bottleneck size of 2 is not significant in either model, and the magnitude of the coefficient goes up as the number of channels increases. For the disharmony-trained models, the coefficient is negative, which is expected given that the model was trained only on disharmony and should thus on average assign lower scores to items that do not fit the training pattern. Figures 3 and 4 show how models at different training steps and with different sized output bottlenecks discriminate identical vs harmonious test data.

5.2 *Featural Analysis* While the discriminator score can statistically be verified to differentiate identity and disharmony when factors like speaker identity are controlled for, it is not the case that the discriminator can reliably function as a classifier of identity vs disharmony as many tokens with disharmony score higher than tokens with harmony. As a follow-up, we explored the possibility that there are features in the fully-connected layer that function in a "constraint-like" manner and that can themselves reliably discriminate identity and disharmony. One possibility is that the lack of such features' influence on the final discriminator score could be that its learned weight is not high enough.

However, across all of the models we did not find features that could qualitatively be characterized as purely detecting the abstract properties of identity vs. disharmony. This is also consistent with the notion that the discriminator scores lower-parameter models, whose features should in principle be motivated to encode broader generalizations about the input were worse at reflecting identity vs. disharmony differences. We show the plot of the best-discriminating feature in the FC layer of the 32 channel model in Fig 5.

Thus, there is no evidence, even in the compressed bottleneck conditions, that any output features in the FC yield categorical constraint-like classifications.

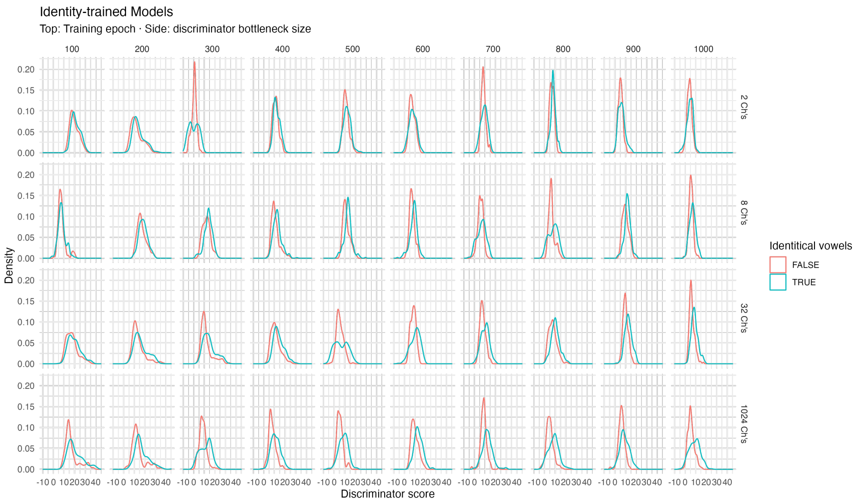

Figure 3: Discriminator score distributions across FC bottleneck sizes of the model and epochs to show the learning dynamics. Each subplot compares a distribution of scores for identical pairs to nonidentical pairs.

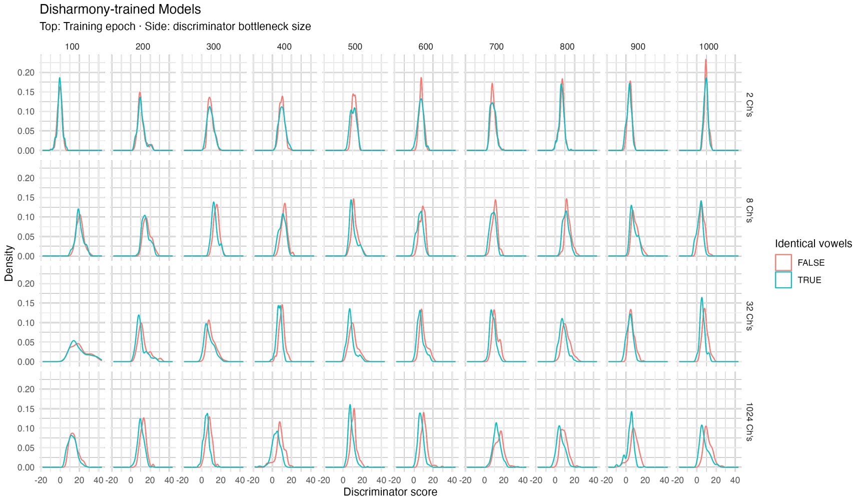

Figure 4: These data show learning for the disharmony pattern. The red curves for nonidentical vowel pairs are shifted slightly rightward, consistent with the notion that the model has some preference for the training-like phonotactic pattern, but there is nevertheless substantial overlap across all conditions.

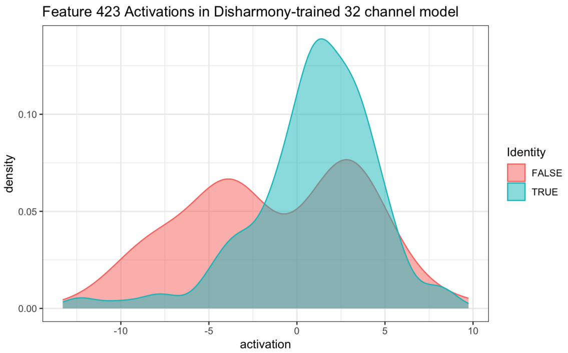

Figure 5: Distributions from activations for feature in the 32-channel model that did the best job of discriminating identical and non-identical vowel pairs with its scores.

Table 1: Effect of harmony across discriminator bottleneck sizes The identityTRUE variable from linear mixed-effects models with Satterthwaite degrees of freedom. Each row represents a separate linear model fit to a particular discriminator's input-output mappings.

Training regime	Bottleneck	Identity β	SE	t	df	p
Identity	2	0.66	0.31	2.10	3.35	0.117
Identity	8	1.74	0.38	4.53	2.83	0.023
Identity	32	2.36	0.48	4.96	3.15	0.014
Identity	1024	3.04	0.34	8.90	11.81	$< 10^{-5}$
Disharmony	2	-0.44	0.37	-1.18	3.13	0.319
Disharmony	8	-2.09	0.43	-4.90	3.09	0.015
Disharmony	32	-2.85	0.33	-8.68	3.68	0.001
Disharmony	1024	-3.77	0.81	-4.68	3.07	0.018

6 Discussion

This study asked whether a WaveGAN discriminator trained under a Wasserstein objective can serve as a model of phonological well-formedness judgments, in the narrow sense that its scalar output assigns higher values to phonotactically "training-like" patterns than to patterns unattested in training. The mixed-effects analyses of discriminator scores indicate that for bottlenecks of 8 channels and above, discriminator scores reliably shift in the predicted direction: identity-trained models assign higher scores to identical-vowel (harmonic) items, while disharmony-trained models assign higher scores to non-identical-vowel items. That these differences do materialize on held out data suggests some degree of generalization beyond memorization of individual word tokens. At the same time, distributions of scores for identity and disharmony overlap substantially as in Figs 3 and 4. From the perspective of weighted-constraint approaches, this at best resembles a situation where an aggregate harmony score reflects the contributions of many weakly informative constraints rather than a small set of high-weight constraints (Smolensky, 1986; Hayes & Wilson, 2008).

A motivation for introducing bottlenecks was that fewer FC features might encourage more interpretable, constraint-like representations—either by forcing compression of information into fewer dimensions or by yielding units whose activations align with an abstract identity constraint. Instead, the smallest bottleneck (2 channels) showed weak and non-significant effects in both training regimes, while larger bottlenecks produced stronger mean shifts. This suggests that, at least in this setting, severe compression impairs the critic's sensitivity to the training distribution, that the model's performance hinges on representing a high degree of phonetic detail in its output layer rather than representing abstract generalizations about the structure of the input. The failure to find single FC units that act as clean identity detectors is compatible with the view that linguistically relevant distinctions are "distributed" or encoded across multiple units and layers rather than localized to a few "constraint" neurons (e.g. Arsenault & Buchsbaum, 2015).

These findings leave open at least three concrete directions for future work. First, future work should test whether similar effects hold on more naturalistic speech with greater speaker and segmental variability, where normalization and exemplar-like structure may interact with phonotactic learning (e.g. Johnson et al., 1997; Pierrehumbert, 2001a; Goldrick & Cole, 2023). Second, the present results suggest that treating the discriminator's scalar output as a proxy for well-formedness may be most informative in controlled comparisons of near-minimal pairs, paralleling approaches in acceptability modeling where scalar scores are interpreted relative to carefully matched alternatives (Lau et al., 2017; Warstadt et al., 2019; de Heer Kloots & Zuidema, 2024). Lastly, the Wasserstein GAN training paradigm provides a template for optimizing or finetuning a model of any architecture whose mapping is from acoustics to a scalar value; for example, existing self-supervised speech models like Wav2Vec2 Baevski et al. (2020) that have been shown to be able to form phoneme-like categories in codebook representations might achieve more "accurate" well-formedness judgments as well as more interpretable constraint-like behavior.

7 Conclusion

Overall, this paper provides some evidence that a WaveGAN discriminator trained with no explicit phonological supervision can develop statistically reliable preferences for the phonotactic pattern present in its training distribution. While these preferences are not consistent in that there is significant overlap between scores for words with vowel disharmony and vowel identity, a Wasserstein discriminator offers a promising starting point for modeling the emergence of phonotactic-like judgments as an unsupervised model mapping acoustic inputs to continuous values.

References

Arsenault, Jessica S & Bradley R Buchsbaum (2015). Distributed neural representations of phonological features during speech perception. *Journal of Neuroscience* 35:2, 634–642.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed & Michael Auli (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33, 12449–12460.

Barman, Sneha Ray, Shakuntala Mahanta & Neeraj Kumar Sharma (2024). Unsupervised modeling of vowel harmony using wavegan. *Proc. SpeechProsody 2024*, 200–204.

Beguš, Gašper (2020a). Modeling unsupervised phonetic and phonological learning in generative adversarial phonology. *Proceedings of the Society for Computation in Linguistics 2020*, Association for Computational Linguistics, New York, New York, 38–48.

Beguš, Gašper (2020b). Modeling unsupervised phonetic and phonological learning in generative adversarial phonology. *Proceedings of the Society for Computation in Linguistics 2020*, 38–48.

Beguš, Gašper (2021). Ciwgan and fiwgan: Encoding information in acoustic data to model lexical learning with generative adversarial networks. *Neural Networks* 139, 305–325.

Beguš, Gašper (2022). Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks. *Computer speech & language* 71, p. 101244.

Beguš, Gašper, Alan Zhou & T Christina Zhao (2023). Encoding of speech in convolutional layers and the brain stem based on language experience. *Scientific Reports* 13:1, p. 6480.

Breiss, Canaan & Adam Albright (2022). Cumulative markedness effects and (non-) linearity in phonotactics .

Gauthier, Jon, Canaan Breiss, Matthew K Leonard & Edward F Chang (2025). Emergent morpho-phonological representations in self-supervised speech models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 28055–28074.

Goldrick, Matthew & Jennifer Cole (2023). Advancement of phonetics in the 21st century: Exemplar models of speech production. *Journal of Phonetics* 99, p. 101254.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville & Yoshua Bengio (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)*.

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

de Heer Kloots, Marianne & Willem Zuidema (2024). Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in wav2vec2. 0. *Proc. INTERSPEECH*.

Johnson, Keith et al. (1997). Speech perception without speaker normalization: An exemplar model. *Talker variability in speech processing* 145–165.

Kuznetsova, Alexandra, Per Bruun Brockhoff, Rune Haubo Bojesen Christensen et al. (2015). Package ‘lmertest’. *R package version* 2:0, p. 734.

Lau, Jey Han, Alexander Clark & Shalom Lappin (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* 41:5, 1202–1241.

Mayer, Connor (2025). Reconciling categorical and gradient models of phonotactics. *Society for Computation in Linguistics* 8:1, p. 5.

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

Pierrehumbert, Janet (2001a). Exemplar dynamics: Word frequency, lenition, and contrast. *Frequency and the Emergence of Linguistic Structure/John Benjamins* .

Pierrehumbert, Janet (2001b). Stochastic phonology. *Glot international* 5:6, 195–207.

Prince, Alan (2004). Optimality theory: Constraint interaction in generative grammar. *University, New Brunswick, and University of Colorado* .

Shain, Cory & Micha Elsner (2019). Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 69–85.

Shain, Cory & Micha Elsner (2020). Acquiring language from speech by learning to remember and predict. Fernández, Raquel & Tal Linzen (eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Online, 195–214, URL `https://aclanthology.org/2020.conll-1.15`.

Smolensky, Paul (1986). Information processing in dynamical systems: Foundations of harmony theory. Rumelhart, David E., James L. McClelland & the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, MIT Press, Cambridge, MA, 194–281.

Taigman, Yaniv, Adam Polyak & Lior Wolf (2016). Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200* .

Warstadt, Alex, Amanpreet Singh & Samuel R. Bowman (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7, 625–641.