Analyzing Phonological Surfeit of the Stimulus in Neural Models

Tovly Deutsch

Ling 215

Abstract

I attempt to replicate the phonological patterns displayed in English plural voicing alternations in language models. In general, the degree to which models replicate human judgements can measure their knowledge about phonology. The experiments in this paper provide additional insight because the English voicing alternation is an example of surfeit of the stimulus, a phenomenon where speakers' productions disregard a statistical pattern of their language in favor of a more natural universal phonological pattern. Previous work has shown that language models are extremely attentive to statistical patterns, often at the expense of human cognitive biases. This tendency would suggest that language models will struggle to replicate phenomena that demonstrate surfeit of the stimulus. Thus, the experiments of this paper attempt to investigate this hypothesis. I find that the models tested perform roughly similarly on inferring voicing acceptability ratings for English words and wug words, demonstrating some ability to replicate humans' surfeit of the stimulus. However, this replication may be spurious due to insufficient exposure to the English lexicon and poor performance relative to naive baselines. I further demonstrate that pretraining and adding wug words to phonological training sets may improve replication of surfeit of the stimulus phenomena.

1 Introduction

The lexicons of languages display many phonological patterns that may not be synchronically motivated. For instance, they may occur due to historical sound changes, extensive borrowing, or sheer coincidence. Simultaneously, these patterns may run contrary to universal and natural phonologically tendencies. Such a phenomenon is observed in English noun plural voicing alternations, as described by Becker, Nevins, and Levine (2012). In the English lexicon, monosyllabic nouns ending in a voiced fricative are more likely to be voiced in their plural forms than similar polysyllables. However, a proposed cross-linguistic phonological generalization is that initial syllables are more faithful that non-initial syllables (Alber 2001; Casali 1998), a phenomena seemingly contradicted by this subset of the English lexicon. Becker, Nevins, and Levine (2012) demonstrate, however, that English speakers do adhere to the cross-linguistic generalization when presented with wug words. This lack of generalization between the lexicon and new productions is termed "surfeit of the stimulus" (Becker, Ketrez, and Nevins 2011).

Adhering to a surfeit of a stimulus phenomena requires some linguistically-universal cognitive bias. Simultaneously, it require the ability to not adhere rigidly to the statistical patterns

of stimuli. Thus, this phenomenon appears one that would be difficult for a statistical model to learn. One such model is a language model which is trained to predict sequences of words. They have also been used as phontactic learners by Mayer and Nelson (2020) in order to judge the acceptability of phonetic sequences in a language. These learners typically adhere to the statistical tendencies of their input data, sometimes in opposition to human behavior that is ingrained or learned from little to no examples (Li et al. 2016; Ponti et al. 2019). Given these tendencies, in this paper I perform a set of experiments to measure how well these models can exhibit the surfeit of the stimulus phenomena demonstrated in English noun pluralization.[1] Some of these experiments mimic those done by Becker, Nevins, and Levine (2012) but with language models as subjects rather than humans. I find that the models surprisingly perform similarly in predicting English and wug word voicing acceptability, although still substantially worse than human raters and similar to naive baselines. Additionally, I find that pretraining and introducing wug words in model training stages may help them ingrain otherwise difficult to learn phonological universals.

2 Methodology

Each experiment in this paper involves training and evaluating language models with English voicing alternating nouns and wug words. Traditionally, language models are trained on sequences of words with no access to constituents of the words like segments or syllables. However, language models are a generalization of models that operate on generic sequences. Thus, they can be used to process or produce sequences of phonological segments as was done by Mayer and Nelson (2020).

2.1 Corpora

The datasets used in this paper originate from two sources: the CMU Pronouncing Dictionary and Becker, Nevins, and Levine (2012). The CMU Pronouncing Dictionary contains pairs of orthographic words and their IPA transcribed forms, sometimes including multiple phonetic forms. This phonetic transcription includes vowel length, primary stress, and secondary stress. Primary stress, where indicated, is included in some experiments; vowel length and secondary stress is removed for all experiments.

Experiments 1 and 2 in Becker, Nevins, and Levine (2012) include two corpora. The first experiment contains 126 English singular nouns ending in either [f] or [θ]. These words are provided in orthographic form so I look up the phonetic form via the CMU Pronouncing Dictionary. Experiment 2 of Becker, Nevins, and Levine (2012) contained 132 wug words that ended in either [f] or [θ]. These are provided in phonetic form, with primary stress information that is used in some of the experiments in the paper. Becker, Nevins, and Levine (2012) provided each word to human participants who rated acceptability between voiced and voiceless plural versions of the word on a 1 (voiceless) to 7 (voiced) scale. Notably, English words demonstrate greater variation in voicing acceptability ratings, having a standard deviation of 1.60 compared to 0.771 for the wug words. This increased variation should make the English voicing acceptability ratings more difficult to infer than the wug voicing ratings.

[1]The code written for these experiments can be found at https://github.com/TovlyDeutsch/215ProjectPublic.

2.2 Model

The model used in this paper is based on the one described by Mayer and Nelson (2020). At its core, it involves an embedding layer followed by a simple recurrent neural network (RNN) optimized during training by the Adam algorithm (Kingma and Ba 2015). The embedding layer transforms each input phonetic token into a vector which is then fed into the RNN. As the model is based on a recurrent neural network, the model can accept as input sequences of any length. In addition to this core, I add a final linear layer that takes the final vector output of the RNN and produces a singular numeric output that represents a voicing acceptability rating. The hyperparameters of the model were selected via manual trial and observation of modifying the existing hyperparameters of Mayer and Nelson (2020). They consist of minibatches of size 64, an embedding layer of size 24, an RNN hidden dimension of size 64, 4 layers in the RNN, a learning rate of 0.005, and 100 epochs of training.

2.3 Pretraining

In some of the experiments of this paper, the model is optionally pretrained before the fine-tuning step. In this process, the model is first trained like a language model to predict segment sequences on a large generic corpus, in this case the CMU Pronouncing Dictionary. After this pretraining is completed, all of the model weights are frozen, except for the weights in the final linear layer. Then, the model is trained on the smaller training dataset relevant to the task, e.g. a set of English nouns. The goal of this pretraining process is to give the model a generic sense of language as a whole with the hypothesis that this knowledge will the help the model perform on some more data-poor task.

2.4 Evaluation

Mean over multiple runs Each training and evaluation cycle of a model has two sources of stochasticity: the random split between test and training data and the stochastic nature of the initialization and optimization algorithms. To account for this stochasticity, each time a model is trained and evaluated in this paper, this training and evaluation is repeated 1000 times and the results presented are averages over these runs. Ordinarily, such repetition is infeasible for neural networks because of the extremely long training time; however, the datasets used in this paper are small enough that this did not become an issue. Repeated tests also allow for significance tests between the performance figures of different model types, an analysis employed in the experiments of this paper.

Evaluation metric For each evaluation, the metric employed is root mean squared error (RMSE). Squaring the error (the difference between the true and inferred result) has the effect making all error positive and thus isolating its magnitude. Squaring the result also places greater penalties on greater deviations, e.g. a raw increase of error by 1 unit from a larger error will have a greater effect on squared error than that same increase of 1 unit starting from a smaller error. The square root is taken finally in order to bring the units of the error measurement in line with the units originally used for the labels and inferences.

Comparisons against naive mean baselines For each experiment in this paper, the result from a naive mean baseline model is provided. This naive baseline simply takes the mean label of the training data and constantly outputs this mean as its inference during evaluation. This baseline serves as comparison for how well models are gaining knowledge of the specific tasks as opposed to simply mimicking mean values.

3 Experiments

3.1 Experiment 1a: Replicating English Voicing Acceptability Ratings

The first experiment attempts to replicate experiment 1 from Becker, Nevins, and Levine (2012), albeit with language model inferences rather than human acceptability judgements. Specifically, the model is fine-tuned on a training subset of 80% of the English voicing alternating nouns given by (Becker, Nevins, and Levine 2012) [2]. These words are labeled with human acceptability judgements of the plural, between 1 (voiceless) and 7 (voiced). After training, the model is evaluated on the remaining 20% of the provided English nouns by tasking the model with inferring voicing acceptability ratings for input words in IPA form.

In addition to this basic paradigm, the model is optionally augmented with pretraining or stress information. For pretraining, the model is first trained as a language model on all words (as segment sequences) in the CMU Pronouncing Dictionary. Then, all weights of the model are frozen except the final linear layer before it is fine-tuned on the English nouns. For the stress modification, the pretraining and fine-tune datasets are augmented with markers of primary stress.

Table 1 displays model evaluation results with various combinations of pretraining and stress inclusion. Table 2 displays significance comparisons between the RMSE of different model types. All of the models performed significantly better than the naive mean baseline, indicating they are learning about the specific task of voicing acceptability inference to some degree. However, this improvement over the baseline is quite small, ranging from 0.01 to 0.04. Additionally, many of the model types did not vary significantly with respect to one another indicating that pretraining and stress inclusion may not be especially useful for this task.

Pretraining	Primary stress	RMSE
✗	✓	1.60
✗	✗	1.60
✓	✓	1.61
✓	✗	1.63
Naive mean baseline		1.64

Table 1: Model evaluations for experiment 1a on predicting English voicing acceptability

[2] The orthographic forms and voicing ratings can be found in Appendix A of the paper.

	Naive mean baseline	Pretraining	Stress	Both
Pretraining	✓			
Stress	✓	✓		
Both	✓	✗	✗	
Neither	✓	✓	✗	✗

Table 2: Pairwise significance in difference between model RSME for experiment 1a. Checkmarks indicate significance in a Tukey HSD test ($p < 0.05, n = 1000$)

3.2 Experiment 1b: Augmented training for Replicating English Voicing Acceptability Ratings

One possible explanation for the poor improvements over baseline seen in experiment 1a is a lack of data. The dataset provided by Becker, Nevins, and Levine (2012) contained only 126 items, only 80% of which are used in training. To expand the dataset size, I extracted additional voicing alternating nouns from the CMU pronouncing dictionary. I selected nouns that ended in [f] or [θ] and had a plural form with only one listed pronunciation. Based on whether this plural form had a voiced ending or not, I assigned it a voicing acceptability rating of 1 or 7. This binary rating is crude and unlikely to be produced by speakers but on average should be strongly correlated with their ratings. These additional examples amounted to 70 new items in the dataset.

Table 3 displays the results from this experiment with the expanded dataset. The raw losses have increased slightly, which is unsurprising given the imprecision of the newly added examples. More interestingly, however, is the greater improvement over the naive baseline, ranging from 0.03 to 0.08. This greater improvement implies that the additional data allows the model to gain a better understanding of the information needed to make voicing acceptability judgements. As with experiment 1a, pretraining and stress addition seem to be largely ineffective in improving performance; the significance comparisons for this experiment are shown in table 4.

Pretraining	Primary stress	RMSE
✓	✓	1.68
✗	✓	1.69
✗	✗	1.69
✓	✗	1.72
Naive mean baseline		1.75

Table 3: Model evaluations on predicting English voicing acceptability with added examples from the CMU Pronouncing Dictionary (experiment 1b)

Based on the results from experiments 1a and 1b, both pretraining and the application of primary stress seem to be ineffective in improving English voicing acceptability inference. However, the RNN-based models do significantly outperform the naive mean baseline, indicating they are gaining some knowledge (however small) about this specific inference problem.

	Naive mean baseline	Pretraining	Stress	Both
Pretraining	✓			
Stress	✓	✓		
Both	✓	✓	✗	
Neither	✓	✓	✗	✗

Table 4: Pairwise significance in difference between model RSME for experiment 1b. Checkmarks indicate significance in a Tukey HSD test ($p < 0.05, n = 1000$)

3.3 Experiment 2: Wug words

After establishing that neural models had a (limited) ability to recognize English voicing alternations, I investigated whether they could extend this knowledge to the productive phonological patterns seen in English speakers. As Becker, Nevins, and Levine (2012) showed, English speakers tended be equally faithful in monosyllable and iambic wug words, while for real English words they treat monosyllables less faithfully than iambs. Importantly, this asymmetry demonstrates they fail to generalize the pattern seen in English to wug words, instead following a proposed protection of initial syllables. Given that machine learning models are extremely attentive to the statistical generalizations of their training data, especially without domain specific modifications, I hypothesized that these models would struggle, trained on English nouns, would struggle to replicate this asymmetry and thus perform more poorly on wug word evaluations.

Table 5 displays the results from this experiment of training on English voicing alternating nouns and evaluating on the wug words provided in Becker, Nevins, and Levine (2012). The training data includes the additional nouns from the CMU Pronouncing Dictionary as described in section 3.2. The naive mean baseline has worsened in performance because the training set (English plurals) and test set (wug words) have a greater difference in means than in experiments 1a and 1b. The non-pretrained models perform similarly to the results seen in 1b. This may indicate the models are failing to learn the generalization of greater impact in English monosyllables, contrary to the hypothesis. However, given that these models are especially attentive to statistical patterns, this failure to generalize may be a result of lack of training data rather than the model being completely unable to capture this pattern. Thus, it is difficult to determine from this experiment alone if the model is successfully honing in on the phonologically natural protection of monosyllables or is simply deprived of enough English stimulus to learn the unnatural pattern in the lexicon.

Unlike in experiments 1a and 1b, adding pretraining results in a significant improvement in performance, as shown in table 6. This may occur because the pretraining may desensitize the model to any unnatural patterns it attempts to learn in the fine-tuning step. Similarly, in the pretraining stage the model may gain some understanding of the natural tendency to protect monosyllables, if this is at all apparent in the broader English lexicon. The addition of primary stress remains ineffective.

Pretraining	Primary stress	RMSE
✓	✗	1.34
✓	✓	1.45
✗	✓	1.68
✗	✗	1.68
Naive mean baseline		2.10

Table 5: Model evaluations on predicting wug word voicing acceptability trained with added examples from the CMU Pronouncing Dictionary (experiment 2)

	Naive mean baseline	Pretraining	Stress	Both
Pretraining	✓			
Stress	✓	✓		
Both	✓	✓	✓	
Neither	✓	✓	✗	✓

Table 6: Pairwise significance in difference between model RSME for experiment 2. Checkmarks indicate significance in a Tukey HSD test ($p < 0.05, n = 1000$)

3.4 Experiment 3: Adding wug words to the fine-tuning training set

Given that exposure to the English lexicon may reduce models' ability to replicate speakers surfeit of the stimulus, I wanted to investigate the addition of wug words to models' training sets. This approach is dissimilar to how human speakers learn language as they are generally not exposed to wug words. However, for the specific case of a statistical phonological model, introducing phonotactically valid wug words that demonstrate some productive phonological pattern may help it replicate patterns not apparent in or contrary to the lexicon. Thus, this third experiment splits the wug dataset into two halves. One half is added to the training set of English nouns. The other half is held out as an evaluation set.

Unsurprisingly, adding wug training data helps in inferring wug ratings, as shown in table 7. The difference in RMSE between the two models is significant ($p < 0.05, n = 1000$). Interestingly, it seems to help the model learn to judge wug words by more than by simply bringing the training mean more in line with the test set. This can be evidenced by the fact that the model with wug words included in training demonstrate a larger difference to a naive mean baseline than the model without wug words in its training set. Thus, the insertion of wug words into model training sets may be an effective strategy for enforcing surfeit of the stimulus phenomena.

Wugs included in training	RMSE	Naive mean baseline	Improvement over naive baseline
✓	0.972	1.62	0.653
✗	1.69	2.10	0.410

Table 7: Model evaluations on predicting wug word voicing acceptability trained with English words and wug words(experiment 3). The two models here have different results for naive mean baselines because they have different training sets.

4 Conclusion

This paper explored how well neural-network-based models could replicate a surfeit of the stimulus phenomenon. Specifically, I experimented with having RNN-based models undertake experiments that Becker, Nevins, and Levine (2012) used to demonstrate an asymmetry in English and wug word plural voicing. Three experiments were conducted focusing on English training to English evaluation, English training to wug evaluation, and English and wug training to wug evaluation. They revealed that, although these models were hypothesized to poorly predict wug acceptability ratings, they did so comparably to English words. However, this seeming success of the models is tempered by their absolute poor performance relative to mean baselines. This poor performance may be due to a lack of training data, thus future work could explore the effects of training on larger sets of English nouns. Additionally, the experiments revealed that both pretraining and adding wug words to training data can aid models in gaining universal generalizations about languages. Such results demonstrate that despite machine learning models' absorption with statistical pattern extraction, there exist effective techniques to imbue them with human-like linguistic preferences.

References

Alber, Birgit (2001). "Maximizing first positions." In: *Proceedings of HILP 5*. Linguistics in Potsdam, pp. 12.1–19.

Becker, Michael, F. Ketrez, and Andrew Nevins (Mar. 1, 2011). "The Surfeit of the Stimulus: Analytic Biases Filter Lexical Statistics in Turkish Laryngeal Alternations". In: *Language* 87, pp. 84–125. DOI: 10.1353/lan.2011.0016.

Becker, Michael, Andrew Nevins, and Jonathan Levine (2012). "ASYMMETRIES IN GENERALIZING ALTERNATIONS TO AND FROM INITIAL SYLLABLES". In: *Language* 88.2. Publisher: Linguistic Society of America, pp. 231–268. ISSN: 0097-8507. URL: https://www.jstor.org/stable/23251831 (visited on 04/19/2020).

Casali, Roderic F. (1998). *Resolving Hiatus*. Google-Books-ID: vomZSWytsSEC. Taylor & Francis. 252 pp. ISBN: 978-0-8153-3149-0.

Kingma, D. P. and L. J. Ba (2015). "Adam: A Method for Stochastic Optimization". In: International Conference on Learning Representations (ICLR). Publisher: Ithaca, NYarXiv.org. URL: https://dare.uva.nl/search?identifier=a20791d3-1aff-464a-8544-268383c33a75 (visited on 04/19/2020).

Li, Jiwei et al. (Nov. 2016). "Deep Reinforcement Learning for Dialogue Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2016. Austin, Texas: Association for Computational Linguistics, pp. 1192–1202. DOI: 10.18653/v1/D16-1127. URL: https://www.aclweb.org/anthology/D16-1127 (visited on 04/20/2020).

Mayer, Connor and Max Nelson (2020). "Phonotactic learning with neural language models". In: *Proceedings of the Society for Computation in Linguistics*. DOI: lingbuzz/004834.

Ponti, Edoardo Maria et al. (Nov. 2019). "Towards Zero-shot Language Modeling". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, pp. 2900–2910. DOI: 10.18653/v1/D19-1288. URL: https://www.aclweb.org/anthology/D19-1288 (visited on 04/20/2020).