Identity Bias & Generalization in a Variable-free Model of Phonotactics

Brandon Prickett
University of Massachusetts Amherst
LSA Annual Meeting
January 3rd, 2020

Road Map

1. **Introduction**
   a) Variables in phonological rules
   b) Variables in phonological constraints
   c) Apparent evidence for variables
   d) Variable-free phonotactic models

Road Map

1. **Introduction**
   a) Variables in phonological rules
   b) Variables in phonological constraints
   c) Apparent evidence for variables
   d) Variable-free phonotactic models
2. **Probabilistic Feature Attention**
   a) Feature attention
   b) Probabilistic feature attention
   c) Learning with PFA
   d) Ambiguous constraint violations
   e) An example with PFA

Road Map

1. **Introduction**
   a) Variables in phonological rules
   b) Variables in phonological constraints
   c) Apparent evidence for variables
   d) Variable-free phonotactic models
2. **Probabilistic Feature Attention**
   a) Feature attention
   b) Probabilistic feature attention
   c) Learning with PFA
   d) Ambiguous constraint violations
   e) An example with PFA
3. **Modeling Identity Bias**
   a) Simulation Set-Up
   b) Results
   c) Why does PFA create an Identity Bias?

Road Map

1. **Introduction**
   a) Variables in phonological rules
   b) Variables in phonological constraints
   c) Apparent evidence for variables
   d) Variable-free phonotactic models
2. **Probabilistic Feature Attention**
   a) Feature attention
   b) Probabilistic feature attention
   c) Learning with PFA
   d) Ambiguous constraint violations
   e) An example with PFA
3. **Modeling Identity Bias**
   a) Simulation Set-Up
   b) Results
   c) Why does PFA create an Identity Bias?
4. **Modeling Identity Generalization**
   a) Simulation Set-Up
   b) Results
   c) Why does PFA cause Identity Generalization?

Road Map

1. **Introduction**
   a) Variables in phonological rules
   b) Variables in phonological constraints
   c) Apparent evidence for variables
   d) Variable-free phonotactic models
2. **Probabilistic Feature Attention**
   a) Feature attention
   b) Probabilistic feature attention
   c) Learning with PFA
   d) Ambiguous constraint violations
   e) An example with PFA
3. **Modeling Identity Bias**
   a) Simulation Set-Up
   b) Results
   c) Why does PFA create an Identity Bias?
4. **Modeling Identity Generalization**
   a) Simulation Set-Up
   b) Results
   c) Why does PFA cause Identity Generalization?
5. **Discussion**
   a) Future work
   b) Conclusions

Introduction

Variables in Phonological Rules

- Halle (1962) first proposed that assimilation and dissimilation patterns could be described using explicit, algebraic variables.

Variables in Phonological Rules

- Halle (1962) first proposed that assimilation and dissimilation patterns could be described using explicit, algebraic variables.
- This provided these kinds of patterns with simpler representations, which is useful since they're typologically common.

Variables in Phonological Rules

- Halle (1962) first proposed that assimilation and dissimilation patterns could be described using explicit, algebraic variables.
- This provided these kinds of patterns with simpler representations, which is useful since they're typologically common.
- For example, if a language has voicing assimilation, this could be represented with the rule:

[-Syllabic] → [αVoice] /_[αVoice]

…where [α] stands for either [+] or [-] and has the same value in both feature bundles.

Variables in Phonological Rules

- Halle (1962) first proposed that assimilation and dissimilation patterns could be described using explicit, algebraic variables.
- This provided these kinds of patterns with simpler representations, which is useful since they're typologically common.
- For example, if a language has voicing assimilation, this could be represented with the rule:

[-Syllabic] → [αVoice] /_[αVoice]

…where [α] stands for either [+] or [-] and has the same value in both feature bundles.

- An analysis of the same pattern that was variable free would require two rules:

[-Syllabic] → [+Voice] /_[+Voice]

[-Syllabic] → [-Voice] /_[-Voice]

Variables in Phonological Constraints

- Constraint-based theories can also use variables in their representations. A constraint like ***[αVoice][-αVoice]** could enforce the same kind of assimilation pattern as Halle's rules.

	*[αVoice][-αVoice]
[td]	*
[dt]	*

Variables in Phonological Constraints

- Constraint-based theories can also use variables in their representations. A constraint like ***[αVoice][-αVoice]** could enforce the same kind of assimilation pattern as Halle's rules.

	*[αVoice][-αVoice]
[td]	*
[dt]	*

- However, recent proposals for phonotactic learning have lacked variables (e.g. Hayes and Wilson 2008; Pater and Moreton 2014). These models require two constraints to represent the same assimilatory process:

	*[+Voice][-Voice]	*[-Voice][+Voice]
[td]		*
[dt]	*	

Apparent Evidence for Variables

- A considerable amount of work has argued against variable-free approaches to phonology (e.g. Moreton 2012; Gallagher 2013; Berent 2013).
  - Whether variables are necessary for any domain of cognition is an active debate in the cognitive science literature (Marcus 2001; Endress et al. 2007; Gervain and Werker 2013; Alhama and Zuidema 2018).

Apparent Evidence for Variables

- A considerable amount of work has argued against variable-free approaches to phonology (e.g. Moreton 2012; Gallagher 2013; Berent 2013).
  - Whether variables are necessary for any domain of cognition is an active debate in the cognitive science literature (Marcus 2001; Endress et al. 2007; Gervain and Werker 2013; Alhama and Zuidema 2018).
- Gallagher (2013) presented two phenomena as evidence for variables:

Apparent Evidence for Variables

- A considerable amount of work has argued against variable-free approaches to phonology (e.g. Moreton 2012; Gallagher 2013; Berent 2013).
  - Whether variables are necessary for any domain of cognition is an active debate in the cognitive science literature (Marcus 2001; Endress et al. 2007; Gervain and Werker 2013; Alhama and Zuidema 2018).
- Gallagher (2013) presented two phenomena as evidence for variables:
  - ***Identity Bias***: identity-based patterns are easier to learn than arbitrary ones

Apparent Evidence for Variables

- A considerable amount of work has argued against variable-free approaches to phonology (e.g. Moreton 2012; Gallagher 2013; Berent 2013).
  - Whether variables are necessary for any domain of cognition is an active debate in the cognitive science literature (Marcus 2001; Endress et al. 2007; Gervain and Werker 2013; Alhama and Zuidema 2018).
- Gallagher (2013) presented two phenomena as evidence for variables:
  - ***Identity Bias***: identity-based patterns are easier to learn than arbitrary ones
  - ***Identity Generalization***: people generalize identity-based patterns in a way that would be predicted by theories that make use of explicit variables (see Berent et al. 2012, Linzen & Gallagher 2017, and Tang & Baer-Henney 2019 for similar results)

Apparent Evidence for Variables

- A considerable amount of work has argued against variable-free approaches to phonology (e.g. Moreton 2012; Gallagher 2013; Berent 2013).
  - Whether variables are necessary for any domain of cognition is an active debate in the cognitive science literature (Marcus 2001; Endress et al. 2007; Gervain and Werker 2013; Alhama and Zuidema 2018).
- Gallagher (2013) presented two phenomena as evidence for variables:
  - ***Identity Bias***: identity-based patterns are easier to learn than arbitrary ones
  - ***Identity Generalization***: people generalize identity-based patterns in a way that would be predicted by theories that make use of explicit variables (see Berent et al. 2012, Linzen & Gallagher 2017, and Tang & Baer-Henney 2019 for similar results)
- While humans demonstrated both of these behaviors in artificial language learning experiments, a variable-free phonotactic learner (Hayes and Wilson 2008) was unable to capture either phenomenon.

Variable-Free Phonotactic Models

- The Hayes and Wilson (2008) phonotactic model learns a probability distribution over all possible words in a language after being trained on that language's lexicon.

Variable-Free Phonotactic Models

- The Hayes and Wilson (2008) phonotactic model learns a probability distribution over all possible words in a language after being trained on that language's lexicon.
  - It represents this probability distribution using a set of weighted constraints like *[+voice] or *[-tense][+word_boundary].

Variable-Free Phonotactic Models

- The Hayes and Wilson (2008) phonotactic model learns a probability distribution over all possible words in a language after being trained on that language's lexicon.
  - It represents this probability distribution using a set of weighted constraints like *[+voice] or *[-tense][+word_boundary].
  - The model's probability estimate for a word is proportional to the weighted sum of that word's constraint violations:

$$p(word_i) = \frac{e^{H_i}}{\sum e^{H_j}} \qquad \textit{where} \qquad H_i = \sum_{c \in C} w_c v_{c,i}$$

Variable-Free Phonotactic Models

- The Hayes and Wilson (2008) phonotactic model learns a probability distribution over all possible words in a language after being trained on that language's lexicon.
  - It represents this probability distribution using a set of weighted constraints like *[+voice] or *[-tense][+word_boundary].
  - The model's probability estimate for a word is proportional to the weighted sum of that word's constraint violations:

$$p(word_i) = \frac{e^{H_i}}{\sum e^{H_j}} \qquad \textit{where} \qquad H_i = \sum_{c \in C} w_c v_{c,i}$$

- The results in this talk are from a different, but similar phonotactic model: GMECCS ("Gradual Maximum Entropy with a Conjunctive Constraint Schema"; Pater & Moreton, 2014; Moreton et al. 2017).

Variable-Free Phonotactic Models

- The Hayes and Wilson (2008) phonotactic model learns a probability distribution over all possible words in a language after being trained on that language's lexicon.
  - It represents this probability distribution using a set of weighted constraints like *[+voice] or *[-tense][+word_boundary].
  - The model's probability estimate for a word is proportional to the weighted sum of that word's constraint violations:

$$p(word_i) = \frac{e^{H_i}}{\sum e^{H_j}} \qquad \textit{where} \qquad H_i = \sum_{c \in C} w_c v_{c,i}$$

- The results in this talk are from a different, but similar phonotactic model: GMECCS ("Gradual Maximum Entropy with a Conjunctive Constraint Schema"; Pater & Moreton, 2014; Moreton et al. 2017).
  - The main difference between the two models is that while Hayes and Wilson's (2008) learner induces its constraint set, GMECCS begins learning with a constraint set that includes every possible conjunction of the relevant phonological features.

Variable-Free Phonotactic Models

- The Hayes and Wilson (2008) phonotactic model learns a probability distribution over all possible words in a language after being trained on that language's lexicon.
  - It represents this probability distribution using a set of weighted constraints like *[+voice] or *[-tense][+word_boundary].
  - The model's probability estimate for a word is proportional to the weighted sum of that word's constraint violations:

$$p(word_i) = \frac{e^{H_i}}{\sum e^{H_j}} \qquad \textit{where} \qquad H_i = \sum_{c \in C} w_c v_{c,i}$$

- The results in this talk are from a different, but similar phonotactic model: GMECCS ("Gradual Maximum Entropy with a Conjunctive Constraint Schema"; Pater & Moreton, 2014; Moreton et al. 2017).
  - The main difference between the two models is that while Hayes and Wilson's (2008) learner induces its constraint set, GMECCS begins learning with a constraint set that includes every possible conjunction of the relevant phonological features.
  - This means that GMECCS's learning process only consists of finding the optimal set of weights for those constraints.

Probabilistic Feature Attention

Feature Attention

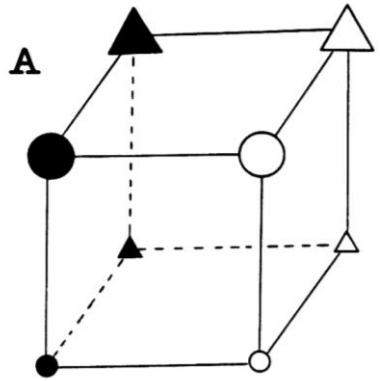
- In cognitive science and machine learning the word “attention” can mean a lot of different things.

Feature Attention

- In cognitive science and machine learning the word “attention” can mean a lot of different things.
- I’m going to be using it in the same sense as Nosofsky (1986), who used “attention” to describe how much a model uses each dimension of a feature space.

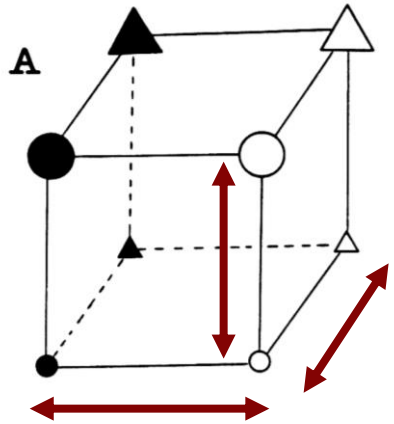
Feature Attention

- In cognitive science and machine learning the word “attention” can mean a lot of different things.
- I’m going to be using it in the same sense as Nosofsky (1986), who used “attention” to describe how much a model uses each dimension of a feature space.
- For example, when learning a visual pattern using shapes of different colors and sizes…



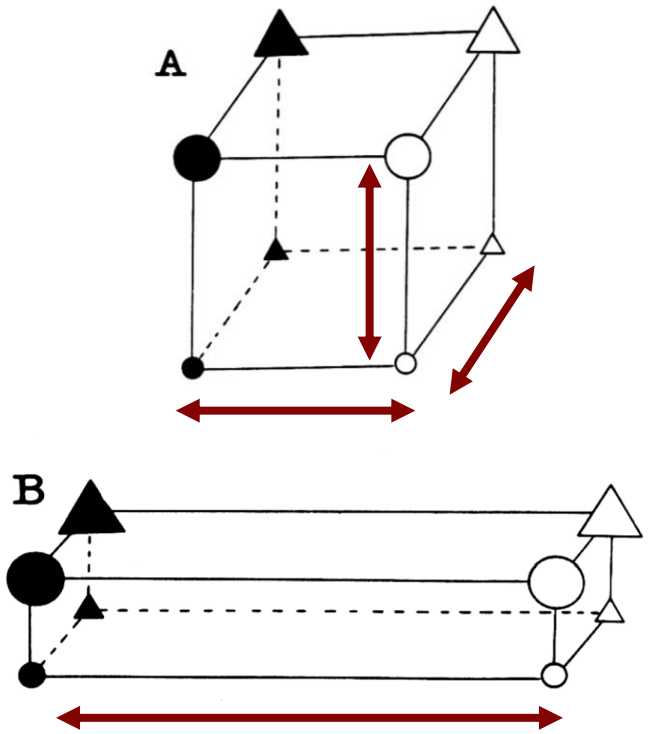
Feature Attention

- In cognitive science and machine learning the word “attention” can mean a lot of different things.
- I’m going to be using it in the same sense as Nosofsky (1986), who used “attention” to describe how much a model uses each dimension of a feature space.
- For example, when learning a visual pattern using shapes of different colors and sizes…
  - One might attend to the features [shape], [size], and [color] equally (as in *A*)…



Feature Attention

- In cognitive science and machine learning the word “attention” can mean a lot of different things.
- I’m going to be using it in the same sense as Nosofsky (1986), who used “attention” to describe how much a model uses each dimension of a feature space.
- For example, when learning a visual pattern using shapes of different colors and sizes…
  - One might attend to the features [shape], [size], and [color] equally (as in *A*)…
  - …or with a disproportionate amount of attention given to [color] (as in *B*).



Probabilistic Feature Attention

- *Probabilistic Feature Attention* (PFA) is a novel mechanism (inspired by Dropout; Srivastava et al. 2014) that randomly distributes attention across different features over the course of learning.

Probabilistic Feature Attention

- *Probabilistic Feature Attention* (PFA) is a novel mechanism (inspired by Dropout; Srivastava et al. 2014) that randomly distributes attention across different features over the course of learning.
- PFA has three main assumptions:

Probabilistic Feature Attention

- *Probabilistic Feature Attention* (PFA) is a novel mechanism (inspired by Dropout; Srivastava et al. 2014) that randomly distributes attention across different features over the course of learning.
- PFA has three main assumptions:

1) **When acquiring phonotactics, a learner doesn't attend to every phonological feature in every word, at every weight update.**

Probabilistic Feature Attention

- *Probabilistic Feature Attention* (PFA) is a novel mechanism (inspired by Dropout; Srivastava et al. 2014) that randomly distributes attention across different features over the course of learning.
- PFA has three main assumptions:

1) **When acquiring phonotactics, a learner doesn't attend to every phonological feature in every word, at every weight update.**
2) **This lack of attention creates ambiguity in the learner's input, since it isn't attending to every feature that might distinguish different data.**

Probabilistic Feature Attention

- *Probabilistic Feature Attention* (PFA) is a novel mechanism (inspired by Dropout; Srivastava et al. 2014) that randomly distributes attention across different features over the course of learning.
- PFA has three main assumptions:

1) **When acquiring phonotactics, a learner doesn't attend to every phonological feature in every word, at every weight update.**
2) **This lack of attention creates ambiguity in the learner's input, since it isn't attending to every feature that might distinguish different data.**
3) **In the face of ambiguity, the grammar errs on the side of assigning constraint violations.**

Learning with PFA

- While PFA could be combined with various learning algorithms and theoretical frameworks, here I pair it with GMECCS (Pater and Moreton 2014; Moreton et al. 2017) and Stochastic Gradient Descent with batch sizes of 1 (all results also hold for batch gradient descent).

Learning with PFA

- While PFA could be combined with various learning algorithms and theoretical frameworks, here I pair it with GMECCS (Pater and Moreton 2014; Moreton et al. 2017) and Stochastic Gradient Descent with batch sizes of 1 (all results also hold for batch gradient descent).
- The learning update for weights in gradient descent is proportional to the difference between the observed violations of a constraint in the training data and the number of violations for that constraint that the model expects to find in the training data, based on its current weights:

$$\Delta w_i = lr \cdot (Obs[v_i] - Exp[v_i])$$

Learning with PFA

- While PFA could be combined with various learning algorithms and theoretical frameworks, here I pair it with GMECCS (Pater and Moreton 2014; Moreton et al. 2017) and Stochastic Gradient Descent with batch sizes of 1 (all results also hold for batch gradient descent).
- The learning update for weights in gradient descent is proportional to the difference between the observed violations of a constraint in the training data and the number of violations for that constraint that the model expects to find in the training data, based on its current weights:

$$\Delta w_i = lr \cdot (Obs[v_i] - \boldsymbol{Exp}[v_i])$$

- PFA introduces ambiguity into the calculation of the model's **expected probabilities**

Learning with PFA

- While PFA could be combined with various learning algorithms and theoretical frameworks, here I pair it with GMECCS (Pater and Moreton 2014; Moreton et al. 2017) and Stochastic Gradient Descent with batch sizes of 1 (all results also hold for batch gradient descent).
- The learning update for weights in gradient descent is proportional to the difference between the observed violations of a constraint in the training data and the number of violations for that constraint that the model expects to find in the training data, based on its current weights:

$$\Delta w_i = lr \cdot (Obs[\boldsymbol{v_i}] - \boldsymbol{Exp}[\boldsymbol{v_i}])$$

- PFA introduces ambiguity into the calculation of the model's **expected probabilities** as well as $\boldsymbol{v_i}$ at each weight update.

Learning with PFA

- While PFA could be combined with various learning algorithms and theoretical frameworks, here I pair it with GMECCS (Pater and Moreton 2014; Moreton et al. 2017) and Stochastic Gradient Descent with batch sizes of 1 (all results also hold for batch gradient descent).
- The learning update for weights in gradient descent is proportional to the difference between the observed violations of a constraint in the training data and the number of violations for that constraint that the model expects to find in the training data, based on its current weights:

$$\Delta w_i = lr \cdot (Obs[\boldsymbol{v_i}] - \boldsymbol{Exp}[\boldsymbol{v_i}])$$

- PFA introduces ambiguity into the calculation of the model's **expected probabilities** as well as $\boldsymbol{v_i}$ at each weight update.
- This added ambiguity means that the weight updates will not always move the model toward a more optimal solution in learning.

Ambiguous Constraint Violations

- For example, let's consider a simplified scenario where words are only one segment long, and there are only four possible segments and two phonological features: [Voice] and [Continuant].

	*[+v]	*[-v]	*[+c]	*[-c]	*[+v, +c]	*[+v, -c]	*[-v, +c]	*[-v, -c]
[d]								
[z]								
[t]								
[s]								

Ambiguous Constraint Violations

- For example, let's consider a simplified scenario where words are only one segment long, and there are only four possible segments and two phonological features: [Voice] and [Continuant].
- When all features are attended to, we get unique violation profiles for every possible segment…

	*[+v]	*[-v]	*[+c]	*[-c]	*[+v, +c]	*[+v, -c]	*[-v, +c]	*[-v, -c]
[d]	1			1		1		
[z]	1		1		1			
[t]		1		1				1
[s]		1	1				1	

Ambiguous Constraint Violations

- For example, let's consider a simplified scenario where words are only one segment long, and there are only four possible segments and two phonological features: [Voice] and [Continuant].
- When all features are attended to, we get unique violation profiles for every possible segment…
- …but if we only attend to [voice], [t]/[s] and [d]/[z] are ambiguous.

	*[+v]	*[-v]	*[+c]	*[-c]	*[+v, +c]	*[+v, -c]	*[-v, +c]	*[-v, -c]
[d] or [z]?								
[d] or [z]?								
[t] or [s]?								
[t] or [s]?								

Ambiguous Constraint Violations

- For example, let's consider a simplified scenario where words are only one segment long, and there are only four possible segments and two phonological features: [Voice] and [Continuant].
- When all features are attended to, we get unique violation profiles for every possible segment…
- …but if we only attend to [voice], [t]/[s] and [d]/[z] are ambiguous.
- And since the grammar errs on the side of assigning constraint violations when faced with ambiguity, each ambiguous datum has a violation profile that's the union of the segments it's ambiguous between.

	*[+v]	*[-v]	*[+c]	*[-c]	*[+v, +c]	*[+v, -c]	*[-v, +c]	*[-v, -c]
[d] or [z]?	1		1	1	1	1		
[d] or [z]?	1		1	1	1	1		
[t] or [s]?		1	1	1			1	1
[t] or [s]?		1	1	1			1	1

An Example with PFA

- Let's say our model is acquiring a language in which only the words [s] and [z] are grammatical.

Attested	**Unattested**
[s], [z]	**[t], [d]**

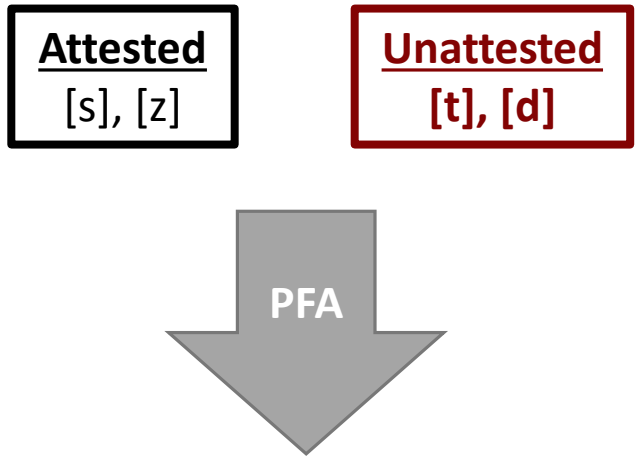
An Example with PFA

- Let's say our model is acquiring a language in which only the words [s] and [z] are grammatical.
  - In a model without PFA, every learning update will add more weight to constraints like *[-Continuant] and less weight to constraints like *[+Continuant].

Attested	**Unattested**
[s], [z]	**[t], [d]**

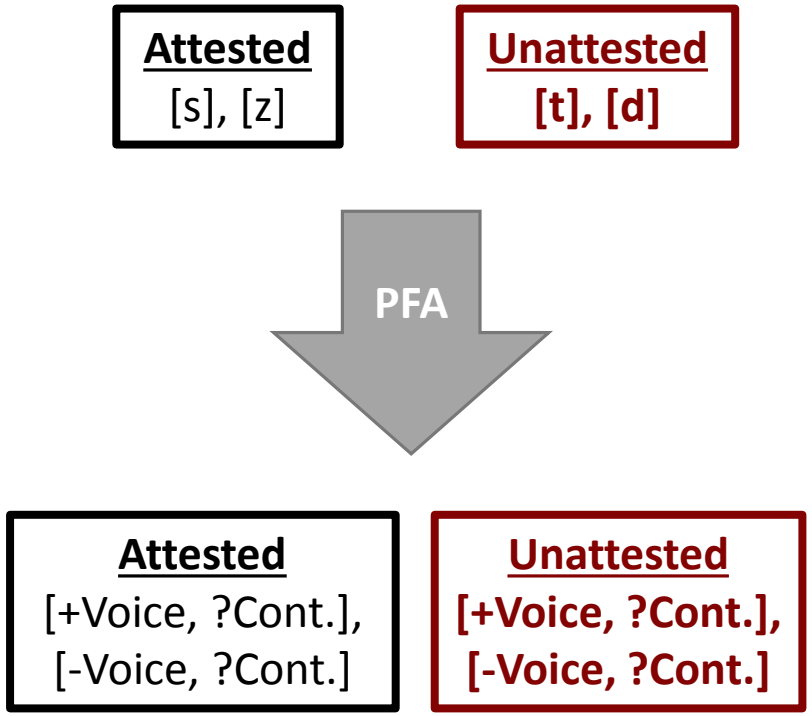
An Example with PFA

- Let's say our model is acquiring a language in which only the words [s] and [z] are grammatical.
  - In a model without PFA, every learning update will add more weight to constraints like *[-Continuant] and less weight to constraints like *[+Continuant].
- However, if PFA is being used, any iteration in which [Continuant] is not attended to will be uninformative for the model…



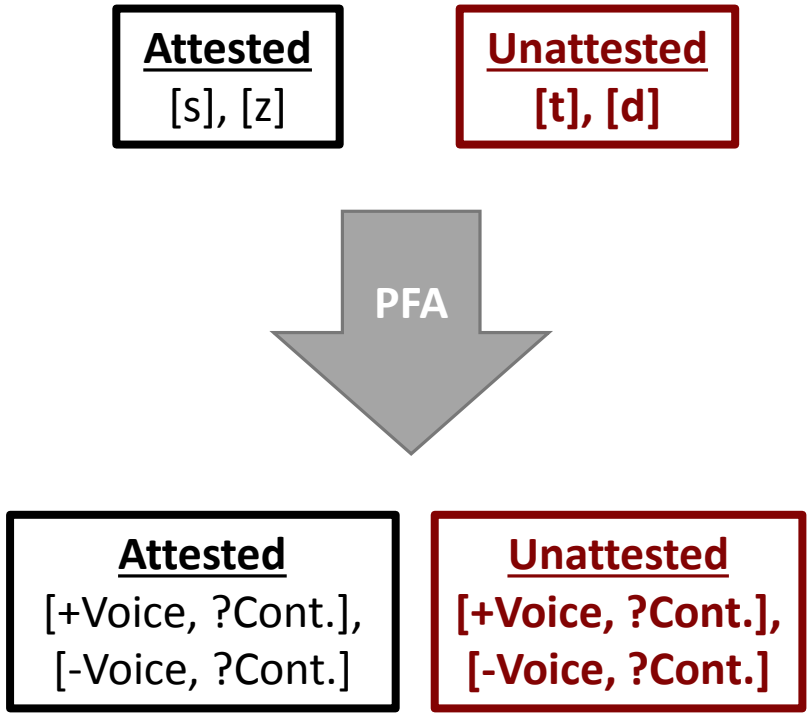
An Example with PFA

- Let's say our model is acquiring a language in which only the words [s] and [z] are grammatical.
  - In a model without PFA, every learning update will add more weight to constraints like *[-Continuant] and less weight to constraints like *[+Continuant].
- However, if PFA is being used, any iteration in which [Continuant] is not attended to will be uninformative for the model…
  - …since attested and unattested sounds will both either be [+Voice, ?Continuant] or [-Voice, ?Continuant].



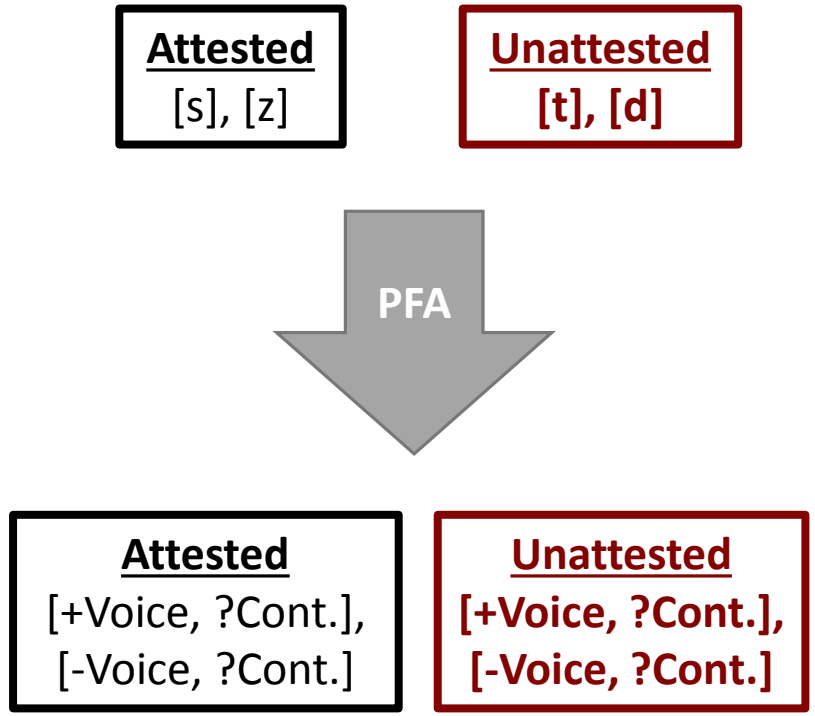
An Example with PFA

- Let's say our model is acquiring a language in which only the words [s] and [z] are grammatical.
  - In a model without PFA, every learning update will add more weight to constraints like *[-Continuant] and less weight to constraints like *[+Continuant].
- However, if PFA is being used, any iteration in which [Continuant] is not attended to will be uninformative for the model…
  - …since attested and unattested sounds will both either be [+Voice, ?Continuant] or [-Voice, ?Continuant].
  - Iterations that *do* attend to [Continuant] will be the only ones in which the model is able to learn anything.



An Example with PFA

- Let's say our model is acquiring a language in which only the words [s] and [z] are grammatical.
  - In a model without PFA, every learning update will add more weight to constraints like *[-Continuant] and less weight to constraints like *[+Continuant].
- However, if PFA is being used, any iteration in which [Continuant] is not attended to will be uninformative for the model…
  - …since attested and unattested sounds will both either be [+Voice, ?Continuant] or [-Voice, ?Continuant].
  - Iterations that *do* attend to [Continuant] will be the only ones in which the model is able to learn anything.
- **In more complex languages, this structured ambiguity can push the model in unexpected directions, causing it to behave differently than models without PFA.**



Modeling Identity Bias

Simulation Set-Up

- To see if PFA allows a model to capture ***Identity Bias***, I simulated Gallagher's (2013) first experiment:

Simulation Set-Up

- To see if PFA allows a model to capture ***Identity Bias***, I simulated Gallagher's (2013) first experiment:
  - The experiment had two language conditions, both of which taught participants a restriction that didn't allow most [+Voice][+Voice] sequences to surface.

Simulation Set-Up

- To see if PFA allows a model to capture ***Identity Bias***, I simulated Gallagher's (2013) first experiment:
  - The experiment had two language conditions, both of which taught participants a restriction that didn't allow most [+Voice][+Voice] sequences to surface.
  - She found that when exceptions to this restriction were identity-based, the pattern was learned more quickly than when the exceptions were arbitrary.

Simulation Set-Up

- To see if PFA allows a model to capture ***Identity Bias***, I simulated Gallagher's (2013) first experiment:
  - The experiment had two language conditions, both of which taught participants a restriction that didn't allow most [+Voice][+Voice] sequences to surface.
  - She found that when exceptions to this restriction were identity-based, the pattern was learned more quickly than when the exceptions were arbitrary.
- Training data for the simulations was identical to the experiment, except all vowels were removed and no information about UR's was given to the model (exceptions to the restriction are in **maroon**):

Simulation Set-Up

- To see if PFA allows a model to capture ***Identity Bias***, I simulated Gallagher's (2013) first experiment:
  - The experiment had two language conditions, both of which taught participants a restriction that didn't allow most [+Voice][+Voice] sequences to surface.
  - She found that when exceptions to this restriction were identity-based, the pattern was learned more quickly than when the exceptions were arbitrary.
- Training data for the simulations was identical to the experiment, except all vowels were removed and no information about UR's was given to the model (exceptions to the restriction are in **maroon**):
  - ***Ident Language***: **[bb]**, [dp], [gp], **[dd]**, [dk], [bt], **[gg]**, [gt], and [bk].

Simulation Set-Up

- To see if PFA allows a model to capture ***Identity Bias***, I simulated Gallagher's (2013) first experiment:
  - The experiment had two language conditions, both of which taught participants a restriction that didn't allow most [+Voice][+Voice] sequences to surface.
  - She found that when exceptions to this restriction were identity-based, the pattern was learned more quickly than when the exceptions were arbitrary.
- Training data for the simulations was identical to the experiment, except all vowels were removed and no information about UR's was given to the model (exceptions to the restriction are in **maroon**):
  - ***Ident Language***: **[bb]**, [dp], [gp], **[dd]**, [dk], [bt], **[gg]**, [gt], and [bk].
  - ***Arbitrary Language***: **[bd]**, **[dg]**, **[gb]**, [bp], [bk], [dp], [dt], [gk], and [gt].

Simulation Set-Up

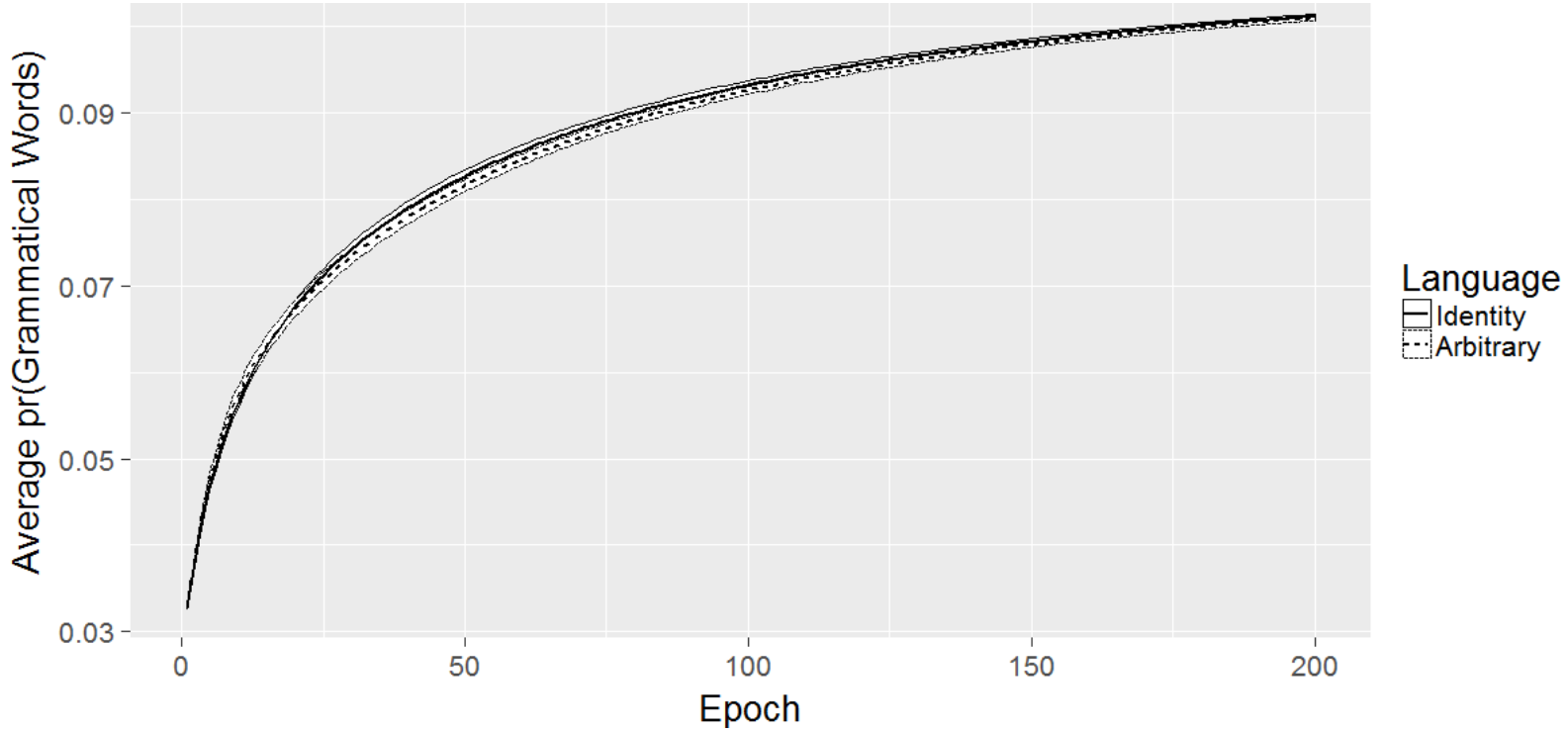
- To see if PFA allows a model to capture ***Identity Bias***, I simulated Gallagher's (2013) first experiment:
  - The experiment had two language conditions, both of which taught participants a restriction that didn't allow most [+Voice][+Voice] sequences to surface.
  - She found that when exceptions to this restriction were identity-based, the pattern was learned more quickly than when the exceptions were arbitrary.
- Training data for the simulations was identical to the experiment, except all vowels were removed and no information about UR's was given to the model (exceptions to the restriction are in **maroon**):
  - ***Ident Language***: **[bb]**, [dp], [gp], **[dd]**, [dk], [bt], **[gg]**, [gt], and [bk].
  - ***Arbitrary Language***: **[bd]**, **[dg]**, **[gb]**, [bp], [bk], [dp], [dt], [gk], and [gt].
- GMECCS was given a constraint set that represented every possible conjunction of the features [±Voice], [±Labial], and [±Dorsal] for ngrams of length 1 and 2.
  - E.g. *[+Voice], *[+Voice, -Labial][-Voice, +Labial], *[+Voice, +Labial, -Dorsal], etc…

Simulation Set-Up

- To see if PFA allows a model to capture ***Identity Bias***, I simulated Gallagher's (2013) first experiment:
  - The experiment had two language conditions, both of which taught participants a restriction that didn't allow most [+Voice][+Voice] sequences to surface.
  - She found that when exceptions to this restriction were identity-based, the pattern was learned more quickly than when the exceptions were arbitrary.
- Training data for the simulations was identical to the experiment, except all vowels were removed and no information about UR's was given to the model (exceptions to the restriction are in **maroon**):
  - ***Ident Language***: **[bb]**, [dp], [gp], **[dd]**, [dk], [bt], **[gg]**, [gt], and [bk].
  - ***Arbitrary Language***: **[bd]**, **[dg]**, **[gb]**, [bp], [bk], [dp], [dt], [gk], and [gt].
- GMECCS was given a constraint set that represented every possible conjunction of the features [±Voice], [±Labial], and [±Dorsal] for ngrams of length 1 and 2.
  - E.g. *[+Voice], *[+Voice, -Labial][-Voice, +Labial], *[+Voice, +Labial, -Dorsal], etc…
- **A model with Identity Bias should learn the Ident Language more quickly than the arbitrary one.**

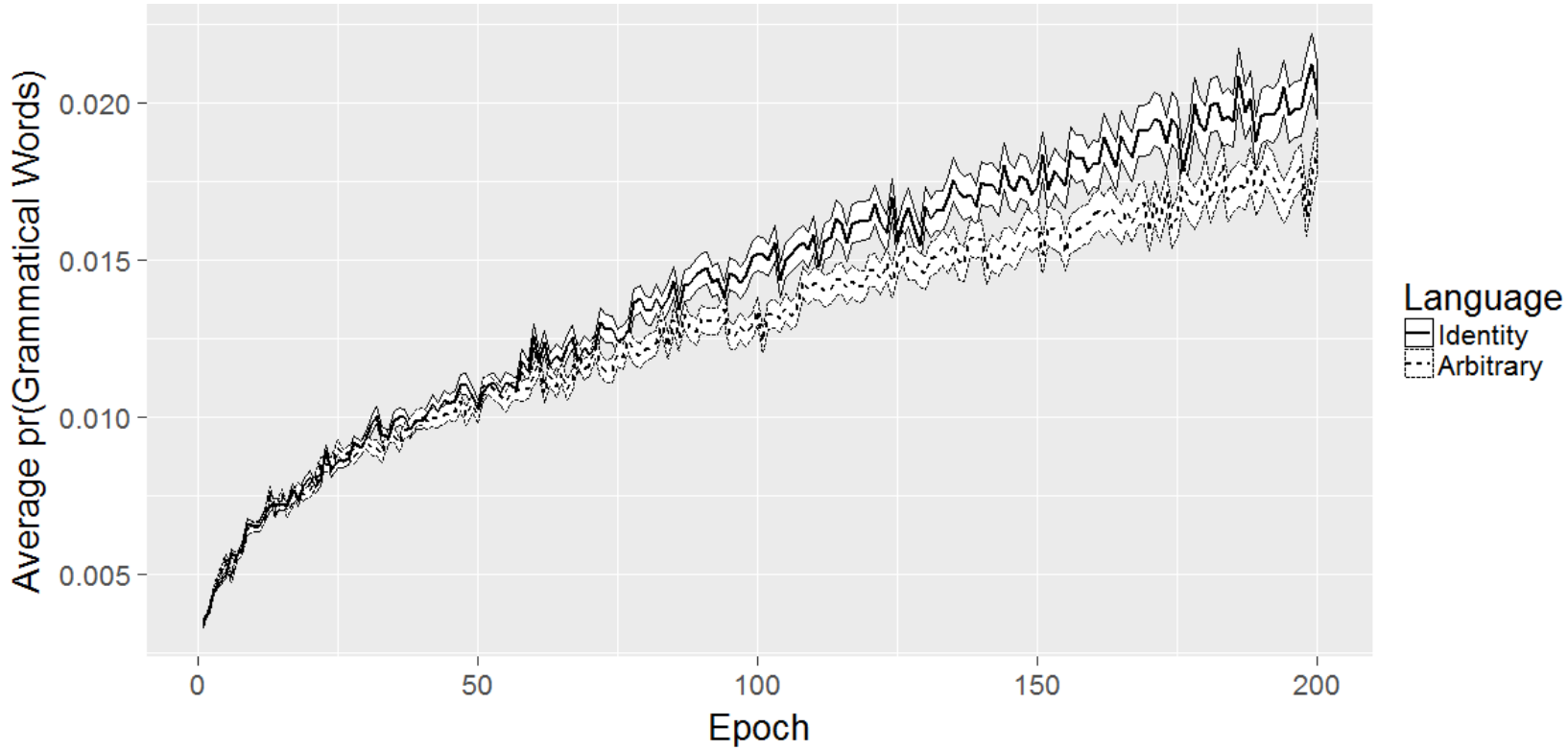
Results (without PFA)

- First, a version of GMECCS with no PFA was trained separately on these two languages for 200 epochs each, with a learning rate of 0.05.



Results (with PFA)

- Then GMECCS was run with PFA, with the same hyperparameters and training data as before, but with only a .25 probability of attending to each feature in any given iteration.



Why does PFA capture Identity Bias?

- The Identity Language is easier for the learner because attested words in that language are more likely to become ambiguous with one another in ways that aid the acquisition of the overall pattern.

Why does PFA capture Identity Bias?

- The Identity Language is easier for the learner because attested words in that language are more likely to become ambiguous with one another in ways that aid the acquisition of the overall pattern.
- For example, [bb] and [dd] are ambiguous with one another at any point in learning where [Labial] is not being attended to.

Why does PFA capture Identity Bias?

- The Identity Language is easier for the learner because attested words in that language are more likely to become ambiguous with one another in ways that aid the acquisition of the overall pattern.
- For example, [bb] and [dd] are ambiguous with one another at any point in learning where [Labial] is not being attended to.
  - This means that any time the model sees either of these data points and isn't paying attention to their labiality, it will move twice as many constraint weights in the correct direction…
  - …Since it will be correctly updating the weights of constraints that [bb] violates *and* constraints that [dd] violates.

Why does PFA capture Identity Bias?

- The Identity Language is easier for the learner because attested words in that language are more likely to become ambiguous with one another in ways that aid the acquisition of the overall pattern.
- For example, [bb] and [dd] are ambiguous with one another at any point in learning where [Labial] is not being attended to.
  - This means that any time the model sees either of these data points and isn't paying attention to their labiality, it will move twice as many constraint weights in the correct direction…
  - …Since it will be correctly updating the weights of constraints that [bb] violates *and* constraints that [dd] violates.
- The Arbitrary Lang's attested words do not have this kind of systematic similarity across data, so the random ambiguity just creates noise in the learning process, making it more difficult to acquire than its counterpart.

Modeling Identity Generalization

Simulation Set-Up

- To test whether humans demonstrated ***Identity Generalization***, Gallagher's (2013) second experiment again trained participants on a voicing restriction with identity-based exceptions.

Simulation Set-Up

- To test whether humans demonstrated ***Identity Generalization***, Gallagher's (2013) second experiment again trained participants on a voicing restriction with identity-based exceptions.
  - However, in this experiment a single pair of identical consonants was withheld from training (e.g. [gg])

Simulation Set-Up

- To test whether humans demonstrated ***Identity Generalization***, Gallagher's (2013) second experiment again trained participants on a voicing restriction with identity-based exceptions.
  - However, in this experiment a single pair of identical consonants was withheld from training (e.g. [gg])
  - In testing, participants were more likely to treat this withheld word as an exception than non-identical words that also violated the restriction (e.g. [dg])

Simulation Set-Up

- To test whether humans demonstrated ***Identity Generalization***, Gallagher's (2013) second experiment again trained participants on a voicing restriction with identity-based exceptions.
  - However, in this experiment a single pair of identical consonants was withheld from training (e.g. [gg])
  - In testing, participants were more likely to treat this withheld word as an exception than non-identical words that also violated the restriction (e.g. [dg])
- The simulation for the Gallagher's (2013) second experiment used the same training data as the Ident Language from before, except with the relevant items withheld:
[bb], [dp], [gp], **[dd]**, [bt], [gt], and [bk].

Simulation Set-Up

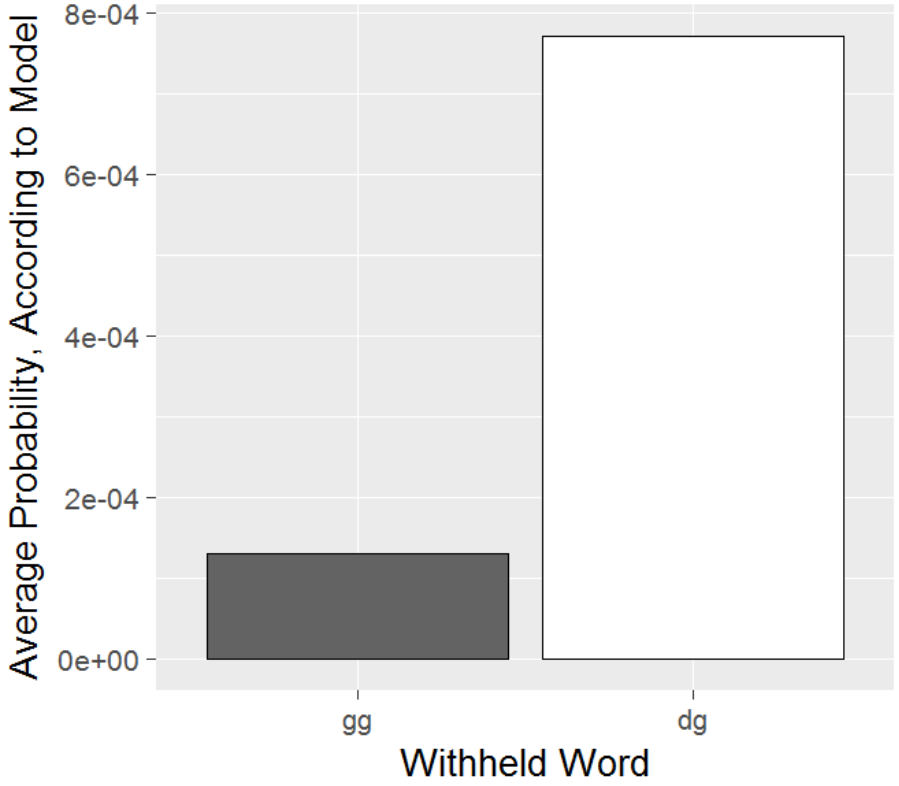
- To test whether humans demonstrated ***Identity Generalization***, Gallagher's (2013) second experiment again trained participants on a voicing restriction with identity-based exceptions.
  - However, in this experiment a single pair of identical consonants was withheld from training (e.g. [gg])
  - In testing, participants were more likely to treat this withheld word as an exception than non-identical words that also violated the restriction (e.g. [dg])
- The simulation for the Gallagher's (2013) second experiment used the same training data as the Ident Language from before, except with the relevant items withheld:
  [bb], [dp], [gp], **[dd]**, [bt], [gt], and [bk].
- And at the end of training, the model was asked to estimate probabilities for the crucial test items: **[gg]** and [dg].

Simulation Set-Up

- To test whether humans demonstrated ***Identity Generalization***, Gallagher's (2013) second experiment again trained participants on a voicing restriction with identity-based exceptions.
  - However, in this experiment a single pair of identical consonants was withheld from training (e.g. [gg])
  - In testing, participants were more likely to treat this withheld word as an exception than non-identical words that also violated the restriction (e.g. [dg])
- The simulation for the Gallagher's (2013) second experiment used the same training data as the Ident Language from before, except with the relevant items withheld:
  [bb], [dp], [gp], **[dd]**, [bt], [gt], and [bk].
- And at the end of training, the model was asked to estimate probabilities for the crucial test items: **[gg]** and [dg].
- **A model exhibiting Identity Generalization should assign more probability to [gg] than to [dg] at the end of training, even though both occur with a probability of 0 in the training data.**

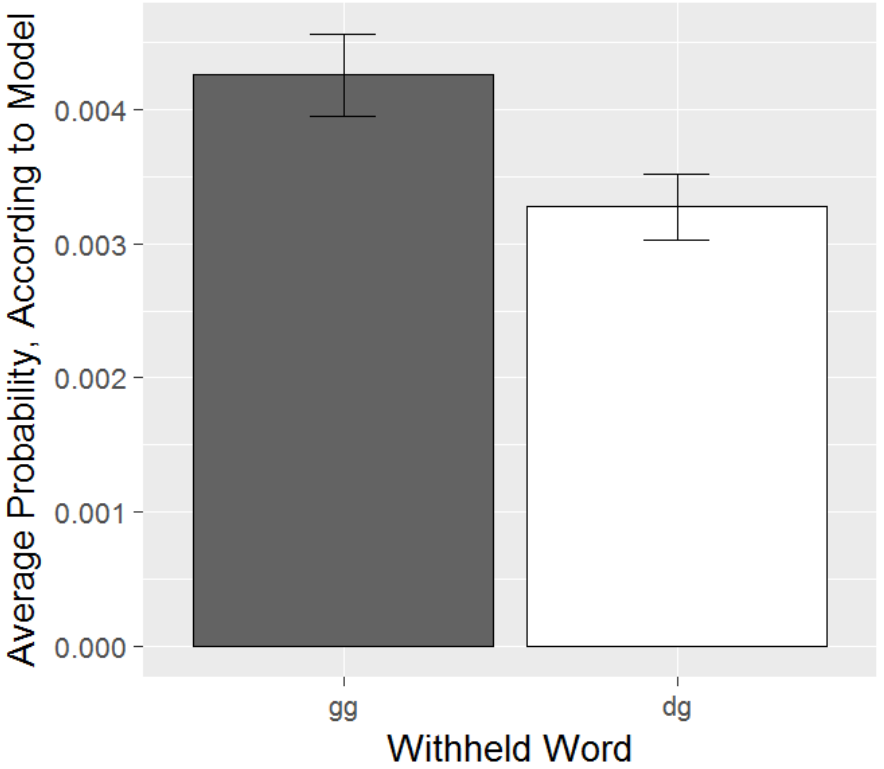
Results (without PFA)

- First, the standard version of GMECCS was trained on the Ident Language for 200 epochs, with a learning rate of 0.05, and tested on the words [gg] and [dg].



Results (with PFA)

- Then GMECCS was run with PFA, with the same hyperparameters and training data as before, but with only a .25 probability of attending to each feature in any given iteration.



Why does PFA capture Identity Generalization?

- This was a result of the fact that, over the course of learning, [dg] is more likely to become ambiguous with other unattested words than [gg] is (because there are more [-Dorsal][+Dorsal] unattested words).

Why does PFA capture Identity Generalization?

- This was a result of the fact that, over the course of learning, [dg] is more likely to become ambiguous with other unattested words than [gg] is (because there are more [-Dorsal][+Dorsal] unattested words).

All Unattested Words, by [Dorsal] Values

[-Dorsal][-Dorsal]		[+Dorsal][-Dorsal]	[-Dorsal][+Dorsal]	[+Dorsal][+Dorsal]
tt	pd	kt	tk	kk
td	pp	kd	tg	kg
tp	pb	kp	dk	gk
tb	bd	kb	***dg***	***gg***
dt	bp	gb	pk	
db		gd	pg	
pt			bg	

Why does PFA capture Identity Generalization?

- This was a result of the fact that, over the course of learning, [dg] is more likely to become ambiguous with other unattested words than [gg] is (because there are more [-Dorsal][+Dorsal] unattested words).

All Unattested Words, by [Dorsal] Values

[-Dorsal][-Dorsal]		[+Dorsal][-Dorsal]	[-Dorsal][+Dorsal]	[+Dorsal][+Dorsal]
tt	pd	kt	**tk**	**kk**
td	pp	kd	**tg**	**kg**
tp	pb	kp	**dk**	**gk**
tb	bd	kb	***dg***	***gg***
dt	bp	gb	**pk**	
db		gd	**pg**	
pt			**bg**	

- Since there are more [-dorsal][+dorsal] words with a probability of 0 than their [+dorsal][+dorsal] counterparts, [dg] is more likely to become ambiguous with other zero-probability words in the training data.

Discussion

Future Work

- One major way to continue the work on PFA is to start applying it to real language data.
  - Berent et al. (2012) showed that variables can capture Identity Generalization demonstrated by native Hebrew speakers.

Future Work

- One major way to continue the work on PFA is to start applying it to real language data.
  - Berent et al. (2012) showed that variables can capture Identity Generalization demonstrated by native Hebrew speakers.
  - To scale up PFA to this kind of data, a more efficient phonotactic learner than GMECCS would likely be needed (e.g. Hayes and Wilson 2008 or Moreton 2019).

Future Work

- One major way to continue the work on PFA is to start applying it to real language data.
  - Berent et al. (2012) showed that variables can capture Identity Generalization demonstrated by native Hebrew speakers.
  - To scale up PFA to this kind of data, a more efficient phonotactic learner than GMECCS would likely be needed (e.g. Hayes and Wilson 2008 or Moreton 2019).
- Are there other phenomena that might be captured using PFA?
  - It's hard to predict what ways PFA will affect a MaxEnt model's learning.

Future Work

- One major way to continue the work on PFA is to start applying it to real language data.
  - Berent et al. (2012) showed that variables can capture Identity Generalization demonstrated by native Hebrew speakers.
  - To scale up PFA to this kind of data, a more efficient phonotactic learner than GMECCS would likely be needed (e.g. Hayes and Wilson 2008 or Moreton 2019).
- Are there other phenomena that might be captured using PFA?
  - It's hard to predict what ways PFA will affect a MaxEnt model's learning.
  - If you're interested in applying PFA to any of the phenomena you're interested in, you can find the software I used at https://github.com/blprickett/Feature-Attention

Conclusions

- Here I’ve shown that two phenomena typically attributed to variables can be captured by a variable-free MaxEnt model equipped with PFA.

Conclusions

- Here I've shown that two phenomena typically attributed to variables can be captured by a variable-free MaxEnt model equipped with PFA.
- In other work that we don't have time to go into, I've found that PFA can model two other phenomena in phonotactic learning:
  - ***Intradimensional Bias***, which Moreton (2012) observed in phonotactic learning and attributed to variables.
  - ***Similarity-based Generalization***, which Cristia et al. (2013) demonstrated in a phonotactic learning experiment and which variables *cannot* capture.

Conclusions

- Here I've shown that two phenomena typically attributed to variables can be captured by a variable-free MaxEnt model equipped with PFA.
- In other work that we don't have time to go into, I've found that PFA can model two other phenomena in phonotactic learning:
  - ***Intradimensional Bias***, which Moreton (2012) observed in phonotactic learning and attributed to variables.
  - ***Similarity-based Generalization***, which Cristia et al. (2013) demonstrated in a phonotactic learning experiment and which variables *cannot* capture.
- Unlike variables, PFA provides a unified account for all of these different phenomena, by assuming structured ambiguity throughout the learning process.

Conclusions

- Here I've shown that two phenomena typically attributed to variables can be captured by a variable-free MaxEnt model equipped with PFA.
- In other work that we don't have time to go into, I've found that PFA can model two other phenomena in phonotactic learning:
  - ***Intradimensional Bias***, which Moreton (2012) observed in phonotactic learning and attributed to variables.
  - ***Similarity-based Generalization***, which Cristia et al. (2013) demonstrated in a phonotactic learning experiment and which variables *cannot* capture.
- Unlike variables, PFA provides a unified account for all of these different phenomena, by assuming structured ambiguity throughout the learning process.
  - This suggests that PFA could be a useful alternative to variables in theories of phonotactic learning.

Acknowledgments

Thanks to the members of the UMass's Sound Workshop, Umass's Phonology Reading Group, the attendees of the 2018 meeting of NECPHON, and the audience at UNC's 2019 Spring Colloquium. For helpful conversations about this work, I also thank Elliott Moreton, Joe Pater, and Gaja Jarosz.

References

Berent, I. (2013). The phonological mind. *Trends in Cognitive Sciences*, *17*(7), 319–327.

Berent, I., Wilson, C., Marcus, G., & Bemis, D. K. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry*, *43*(1), 97–119.

Cristia, A., Mielke, J., Daland, R., & Peperkamp, S. (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, *4*(2), 259–285.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

Gallagher, G. (2013). Learning the identity effect as an artificial language: Bias and generalisation. *Phonology*, *30*(2), 253–295.

Gervain, J., & Werker, J. F. (2013). Learning non-adjacent regularities at age 0; 7. *Journal of Child Language*, *40*(4), 860–872.

Halle, M. (1962). A descriptive convention for treating assimilation and dissimilation. *Quarterly Progress Report*, *66*, 295–296.

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

Linzen, T., & Gallagher, G. (2017). Rapid generalization in phonotactic learning. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1).

Marcus, G. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.

Marcus, G., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Moreton, E. (2012). Inter-and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, *67*(1), 165–183.

Moreton, E. (2019). Constraint breeding during on-line incremental learning. *Proceedings of the Society for Computation in Linguistics*, *2*(1), 69–80.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Tang, K., & Baer-Henney, D. (2019). *Disentangling L1 and L2 effects in artificial language learning*. Presented at the Manchester Phonology Meeting, Manchester, UK. Retrieved from http://www.lel.ed.ac.uk/mfm/27mfm-abbk.pdf

Appendix

Gallagher (2013): Experiment 1 Design

- This experiment was aimed at discovering whether identity-based patterns were easier to learn than more arbitrary ones (i.e. ***Identity Bias***).

Gallagher (2013): Experiment 1 Design

- This experiment was aimed at discovering whether identity-based patterns were easier to learn than more arbitrary ones (i.e. ***Identity Bias***).
- To test this, Gallagher trained participants on one of two devoicing patterns

[babu] → [bapu]
[dabu] → [dapu]
[gabu] → [gapu]

Gallagher (2013): Experiment 1 Design

- This experiment was aimed at discovering whether identity-based patterns were easier to learn than more arbitrary ones (i.e. ***Identity Bias***).
- To test this, Gallagher trained participants on one of two devoicing patterns: a pattern with unsystematic exceptions to devoicing ("Arbitrary Lang" below)

Examples from Arbitrary Lang

Alternating	Exception
[babu] → [bapu]	[badu] → [badu]
[dabu] → [dapu]	[dagu] → [daku]
[gabu] → [gapu]	[gabu] → [gabu]

Gallagher (2013): Experiment 1 Design

- This experiment was aimed at discovering whether identity-based patterns were easier to learn than more arbitrary ones (i.e. ***Identity Bias***).
- To test this, Gallagher trained participants on one of two devoicing patterns: a pattern with unsystematic exceptions to devoicing ("Arbitrary Lang" below) and a pattern in which devoicing did not occur when a word's consonants were identical ("Identity Lang" below).

Examples from Arbitrary Lang

Alternating	Exception
[babu] → [bapu]	[badu] → [badu]
[dabu] → [dapu]	[dagu] → [daku]
[gabu] → [gapu]	[gabu] → [gabu]

Examples from Identity Lang

Alternating
[badu] → [batu]
[dagu] → [daku]
[gadu] → [gatu]

Gallagher (2013): Experiment 1 Design

- This experiment was aimed at discovering whether identity-based patterns were easier to learn than more arbitrary ones (i.e. ***Identity Bias***).
- To test this, Gallagher trained participants on one of two devoicing patterns: a pattern with unsystematic exceptions to devoicing ("Arbitrary Lang" below) and a pattern in which devoicing did not occur when a word's consonants were identical ("Identity Lang" below).

Examples from Arbitrary Lang

Alternating	Exception
[babu] → [bapu]	[badu] → [badu]
[dabu] → [dapu]	[dagu] → [daku]
[gabu] → [gapu]	[gabu] → [gabu]

Examples from Identity Lang

Alternating	Exception
[badu] → [batu]	[babu] → [babu]
[dagu] → [daku]	[dadu] → [dadu]
[gadu] → [gatu]	[gagu] → [gagu]

Gallagher (2013): Experiment 1 Design

- This experiment was aimed at discovering whether identity-based patterns were easier to learn than more arbitrary ones (i.e. ***Identity Bias***).
- To test this, Gallagher trained participants on one of two devoicing patterns: a pattern with unsystematic exceptions to devoicing ("Arbitrary Lang" below) and a pattern in which devoicing did not occur when a word's consonants were identical ("Identity Lang" below).

Examples from Arbitrary Lang

Alternating	Exception
[babu] → [bapu]	[badu] → [badu]
[dabu] → [dapu]	[dagu] → [daku]
[gabu] → [gapu]	[gabu] → [gabu]

Examples from Identity Lang

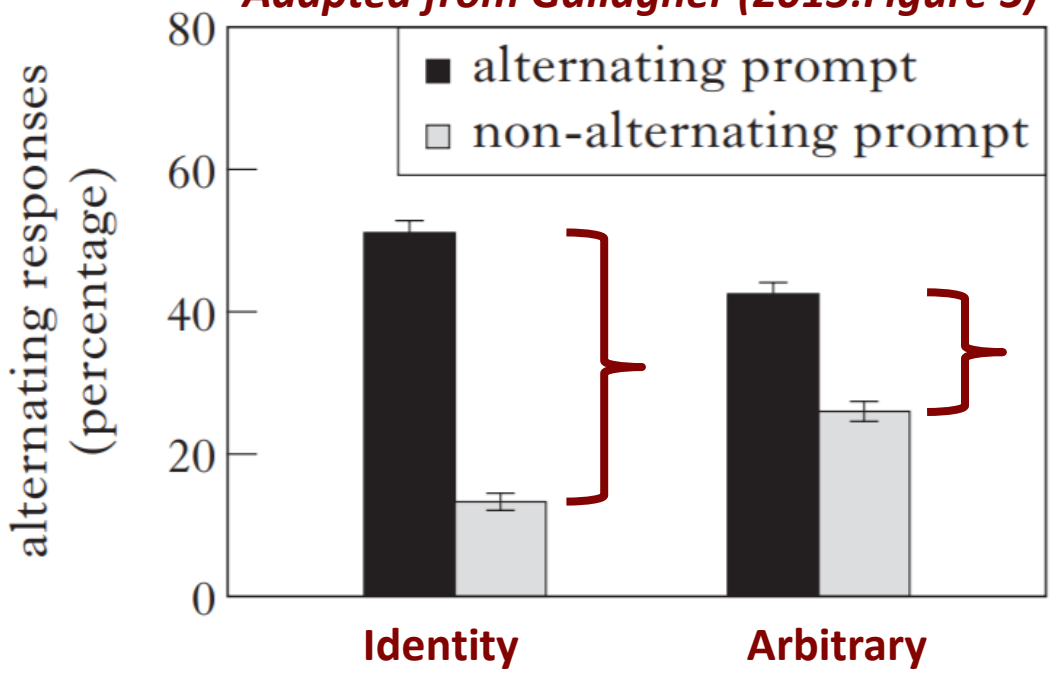
Alternating	Exception
[badu] → [batu]	[babu] → [babu]
[dagu] → [daku]	[dadu] → [dadu]
[gadu] → [gatu]	[gagu] → [gagu]

- Participants were trained on these mappings and then tested on novel mappings that involved the same crucial consonantal pairs.

Gallagher (2013): Experiment 1 Results

- The results for this experiment demonstrated that participants learned the Identity Language better than the Arbitrary Language.

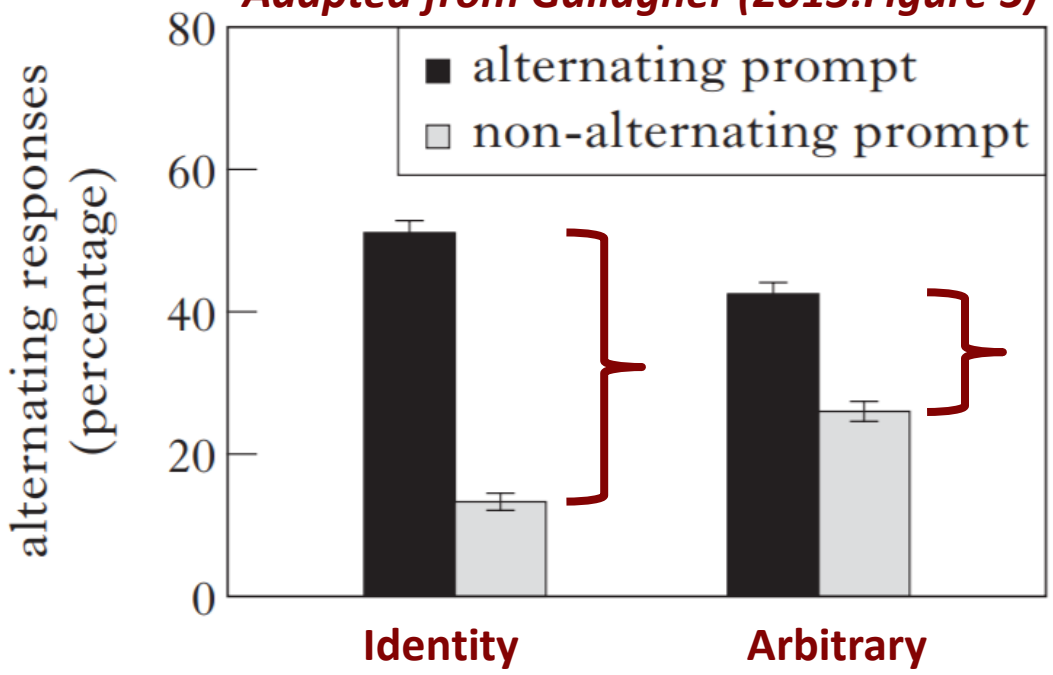
Adapted from Gallagher (2013:Figure 3)



Gallagher (2013): Experiment 1 Results

- The results for this experiment demonstrated that participants learned the Identity Language better than the Arbitrary Language.
- This suggests that an Identity Bias was affecting their learning.

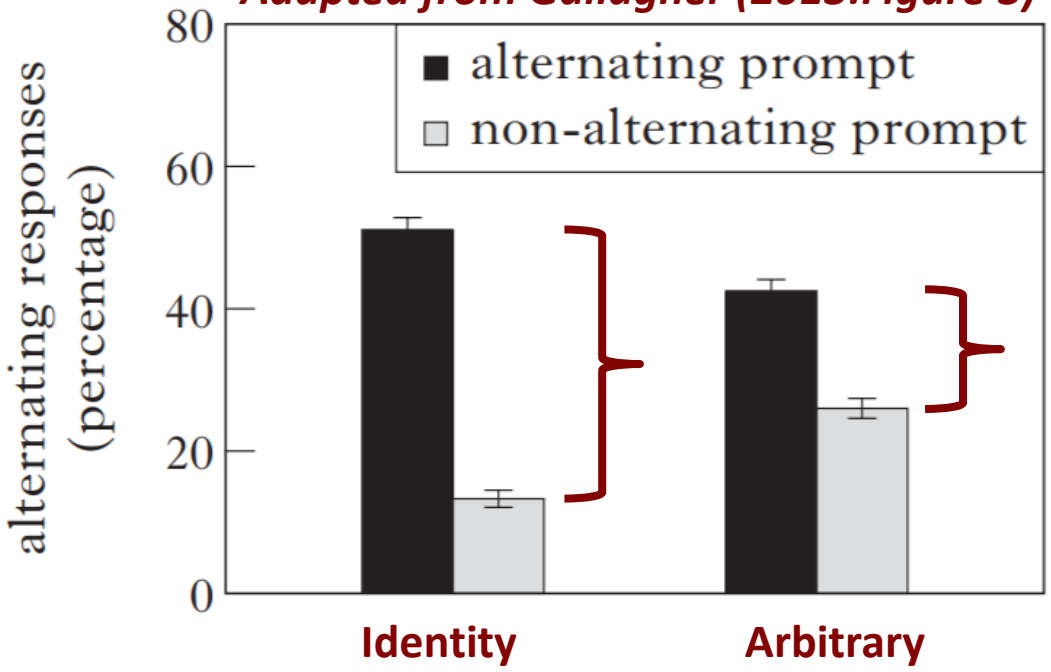
Adapted from Gallagher (2013:Figure 3)



Gallagher (2013): Experiment 1 Results

- The results for this experiment demonstrated that participants learned the Identity Language better than the Arbitrary Language.
- This suggests that an Identity Bias was affecting their learning.
- Gallagher (2013) showed that the Hayes and Wilson (2008) learner could not model this bias unless variables were added.

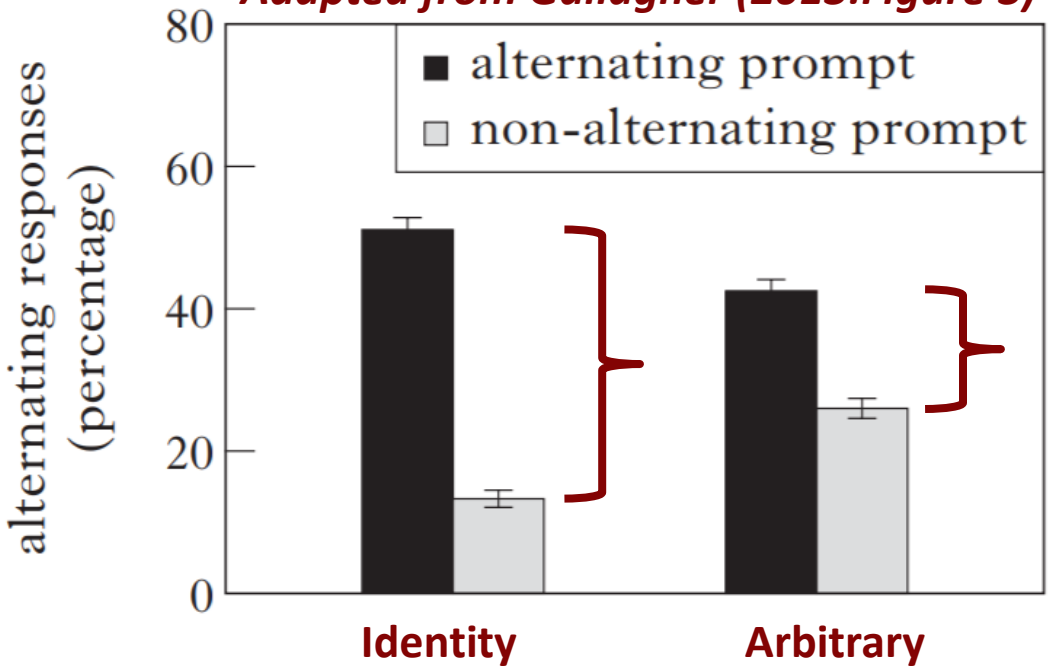
Adapted from Gallagher (2013:Figure 3)



Gallagher (2013): Experiment 1 Results

- The results for this experiment demonstrated that participants learned the Identity Language better than the Arbitrary Language.
- This suggests that an Identity Bias was affecting their learning.
- Gallagher (2013) showed that the Hayes and Wilson (2008) learner could not model this bias unless variables were added.
  - This is because variables cause the identity-based pattern to be structurally simpler than the arbitrary language.

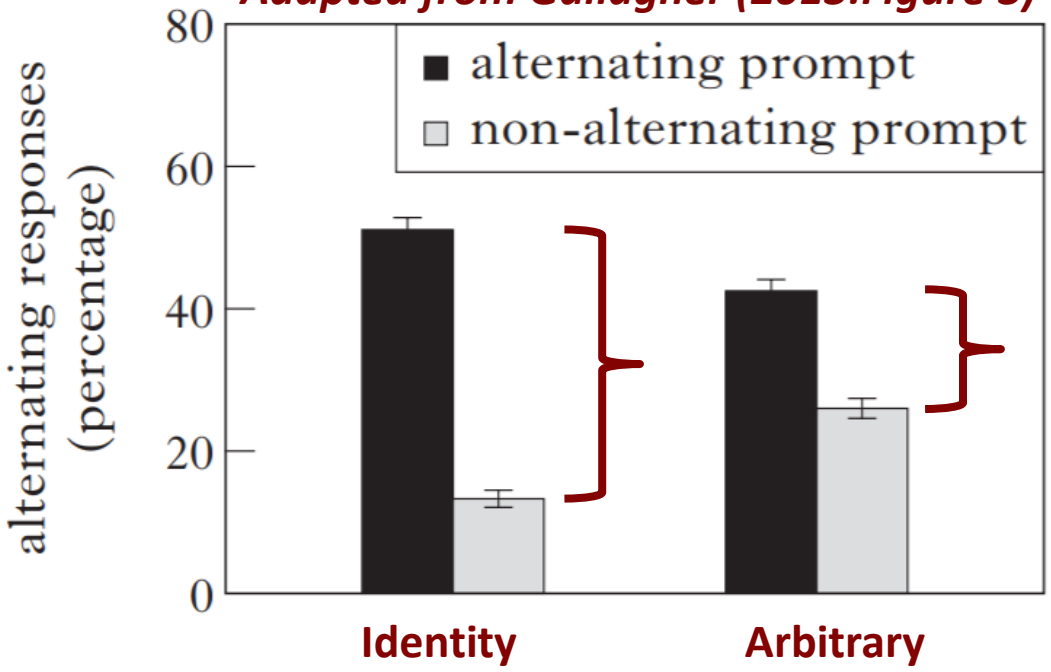
Adapted from Gallagher (2013:Figure 3)



Gallagher (2013): Experiment 1 Results

- The results for this experiment demonstrated that participants learned the Identity Language better than the Arbitrary Language.
- This suggests that an Identity Bias was affecting their learning.
- Gallagher (2013) showed that the Hayes and Wilson (2008) learner could not model this bias unless variables were added.
  - This is because variables cause the identity-based pattern to be structurally simpler than the arbitrary language.
  - Without variables, the two patterns require the same number of constraints to represent.

Adapted from Gallagher (2013:Figure 3)



Gallagher (2013): Experiment 2 Design

- This experiment tested whether participants generalized identity-based patterns to novel segments (i.e. ***Identity Generalization***).

Gallagher (2013): Experiment 2 Design

- This experiment tested whether participants generalized identity-based patterns to novel segments (i.e. ***Identity Generalization***).
- To test this, the Identity Language from Experiment 1 was altered so that a single pair of identical consonants was withheld from training.

Alternating	Exception
[badu] → [batu]	[babu] → [babu]
[dagu] → [daku]	[dadu] → [dadu]

Gallagher (2013): Experiment 2 Design

- This experiment tested whether participants generalized identity-based patterns to novel segments (i.e. ***Identity Generalization***).
- To test this, the Identity Language from Experiment 1 was altered so that a single pair of identical consonants was withheld from training.
- Participants were then tested on this pair, as well as another pair that was not identical.

Alternating	Exception	Withheld
[badu] → [batu]	[babu] → [babu]	[gagu] → ?
[dagu] → [daku]	[dadu] → [dadu]	[gadu] → ?

Gallagher (2013): Experiment 2 Design

- This experiment tested whether participants generalized identity-based patterns to novel segments (i.e. ***Identity Generalization***).
- To test this, the Identity Language from Experiment 1 was altered so that a single pair of identical consonants was withheld from training.
- Participants were then tested on this pair, as well as another pair that was not identical.

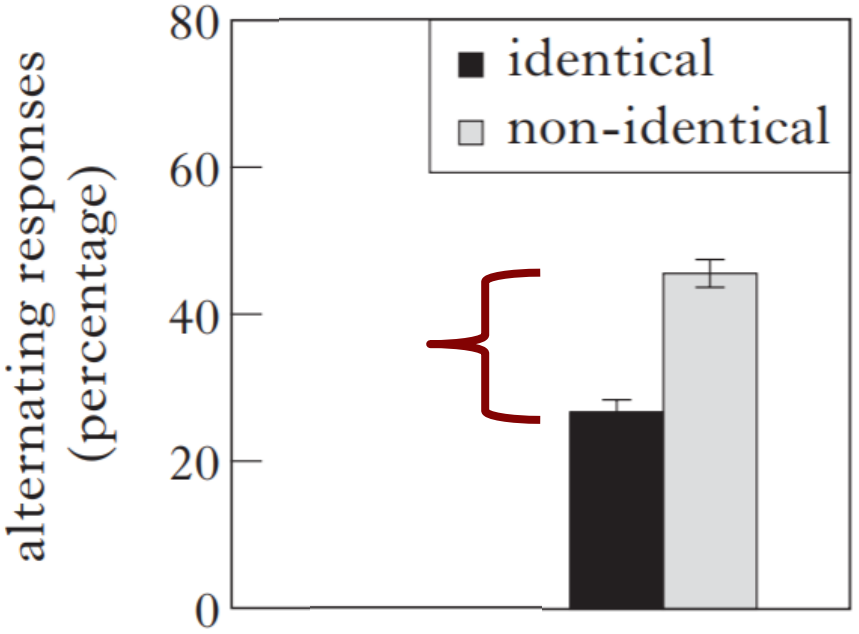
Alternating	Exception	Withheld
[badu] → [batu]	[babu] → [babu]	[gagu] → ?
[dagu] → [daku]	[dadu] → [dadu]	[gadu] → ?

- If participants learned the identity pattern in a generalizable way, the devoicing process should be applied to the non-identical novel pair but not the identical one.

Gallagher (2013): Experiment 2 Results

- The results showed that participants were more likely to devoice non-identical withheld consonant pairs than their counterparts.

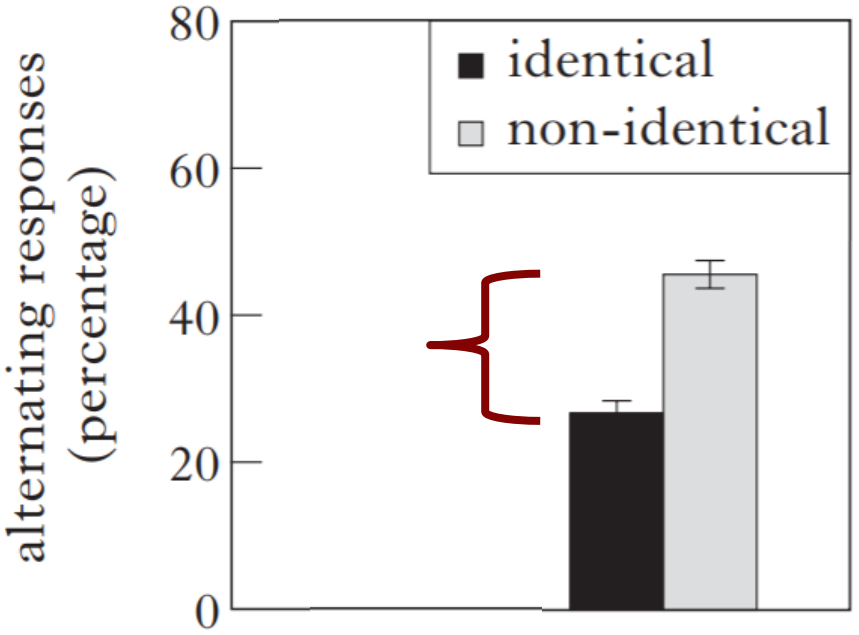
Adapted from Gallagher (2013:Figure 4)



Gallagher (2013): Experiment 2 Results

- The results showed that participants were more likely to devoice non-identical withheld consonant pairs than their counterparts.
- This suggests that they were properly generalizing the identity-based pattern of exceptionality in the language.

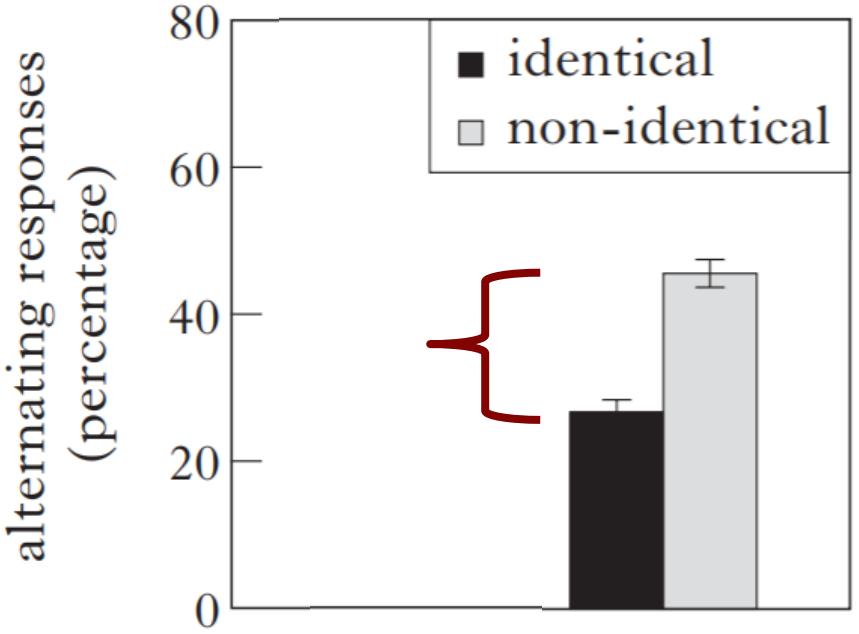
Adapted from Gallagher (2013:Figure 4)



Gallagher (2013): Experiment 2 Results

- The results showed that participants were more likely to devoice non-identical withheld consonant pairs than their counterparts.
- This suggests that they were properly generalizing the identity-based pattern of exceptionality in the language.
- The Hayes and Wilson (2008) model cannot capture this kind of generalization because it doesn't have any way of representing similarity across segments.

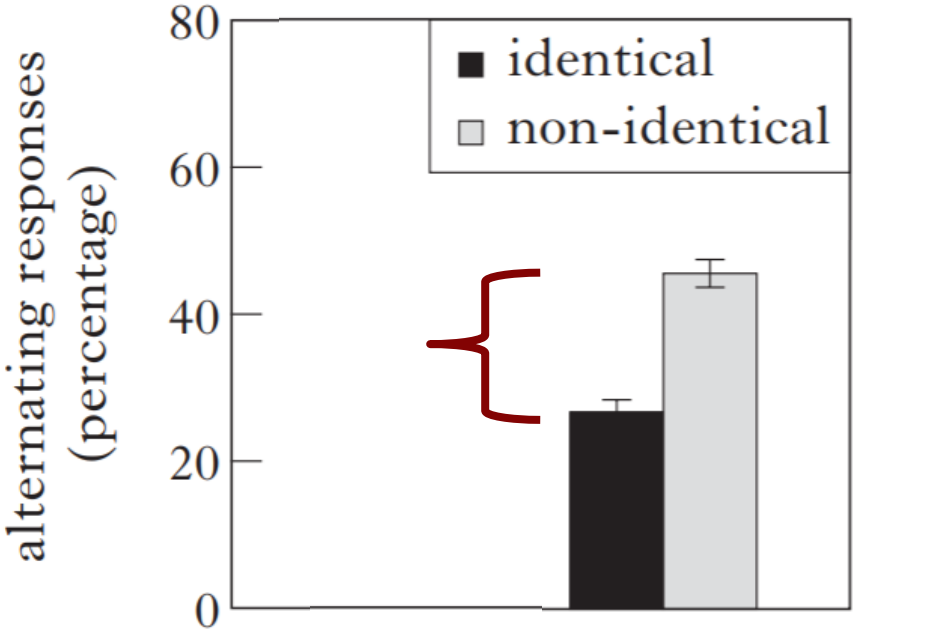
Adapted from Gallagher (2013:Figure 4)



Gallagher (2013): Experiment 2 Results

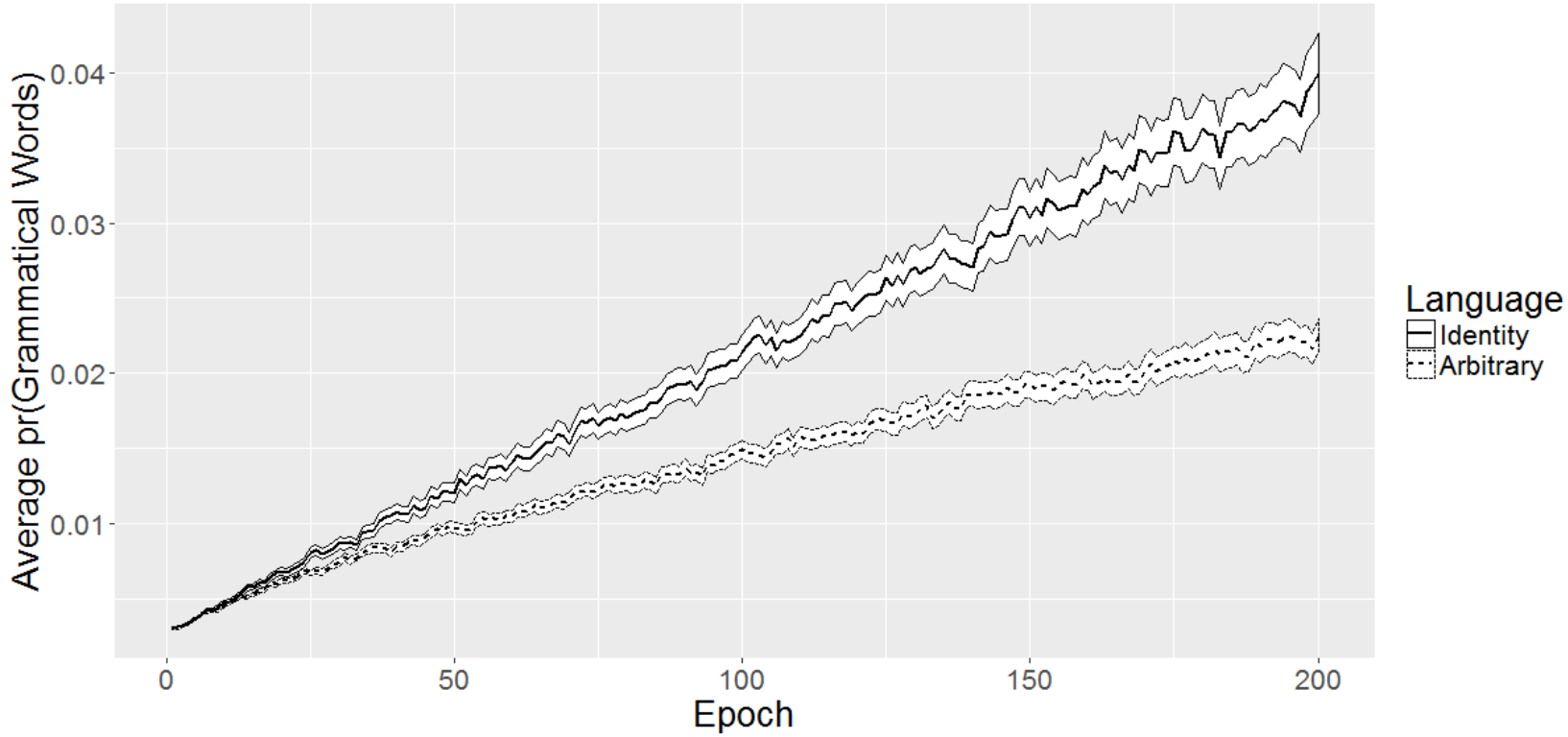
- The results showed that participants were more likely to devoice non-identical withheld consonant pairs than their counterparts.
- This suggests that they were properly generalizing the identity-based pattern of exceptionality in the language.
- The Hayes and Wilson (2008) model cannot capture this kind of generalization because it doesn't have any way of representing similarity across segments.
  - i.e. there's no way for the model to represent that [gagu], [babu], and [dadu] all have something in common with variable-free constraints, so no extra probability will be given to the withheld word.

Adapted from Gallagher (2013:Figure 4)



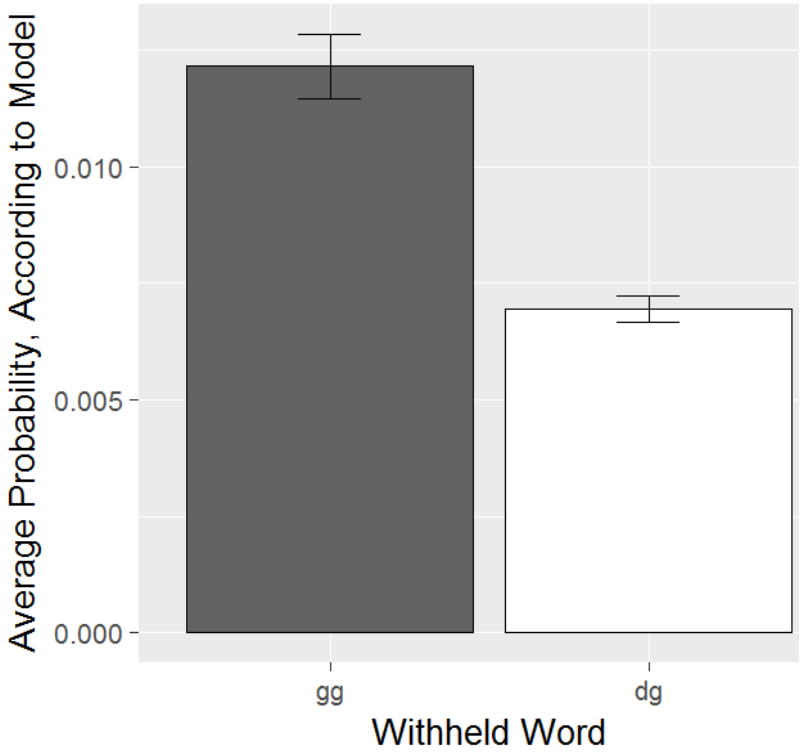
Results (with PFA, only exceptions)

- This effect is exaggerated even more when GMECCS is trained to just differentiate the exceptions to the devoicing pattern from all other words.



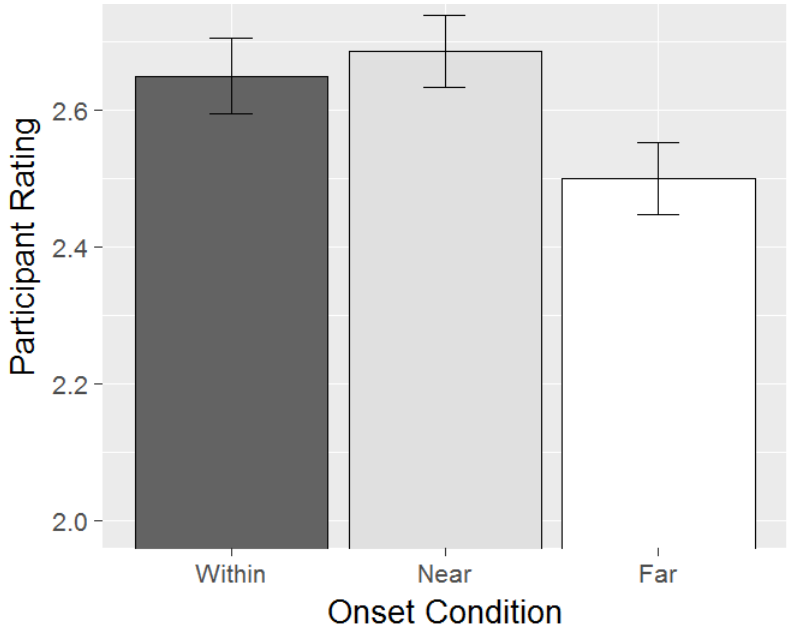
Results (with PFA, only exceptions)

- Again, the effect is even more apparent when GMECCS is trained to just pick out the exceptions to the devoicing pattern.

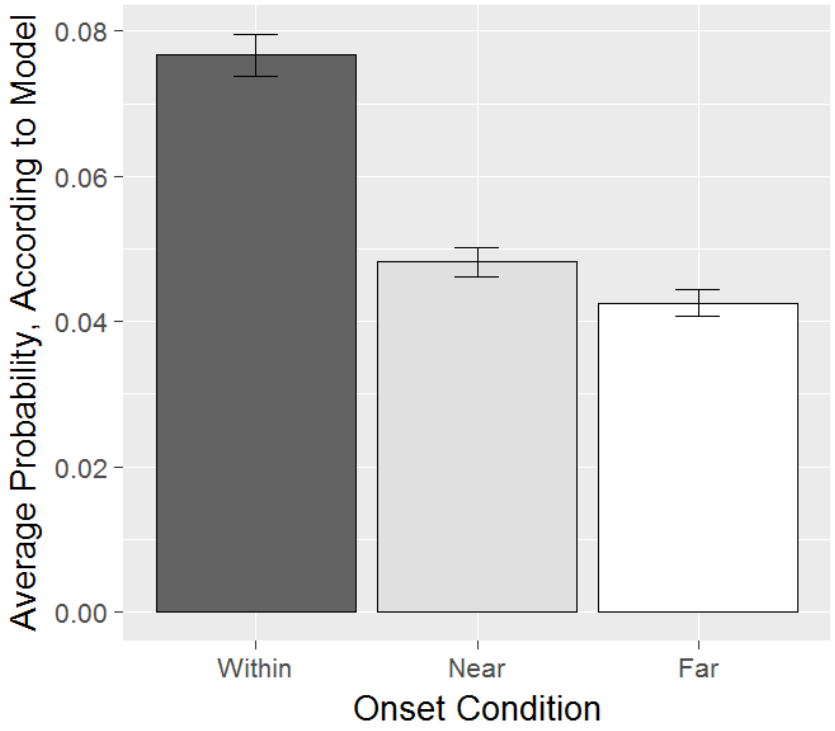


Similarity-Based Generalization (Cristia et al. 2013)

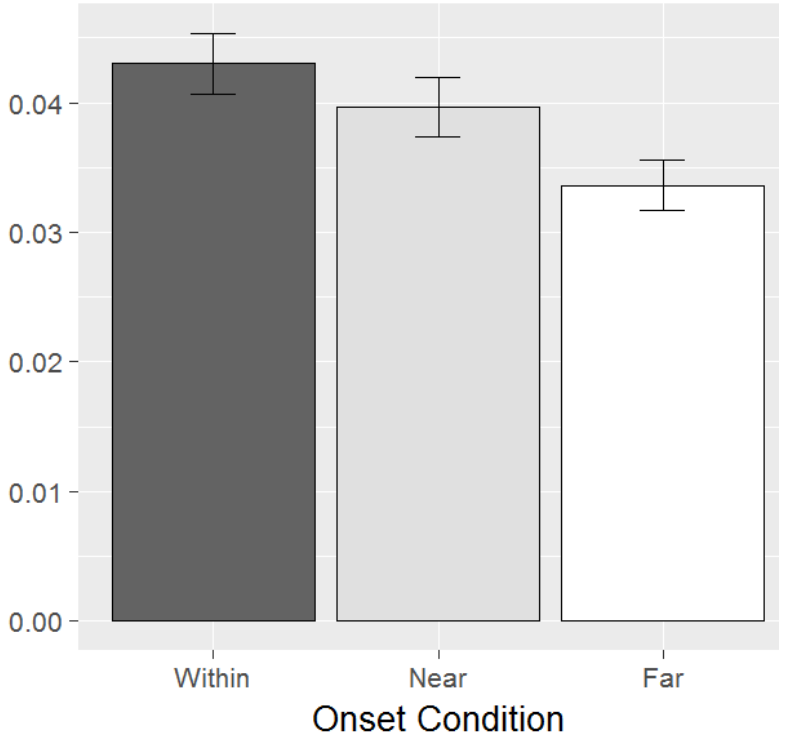
Humans



Without PFA

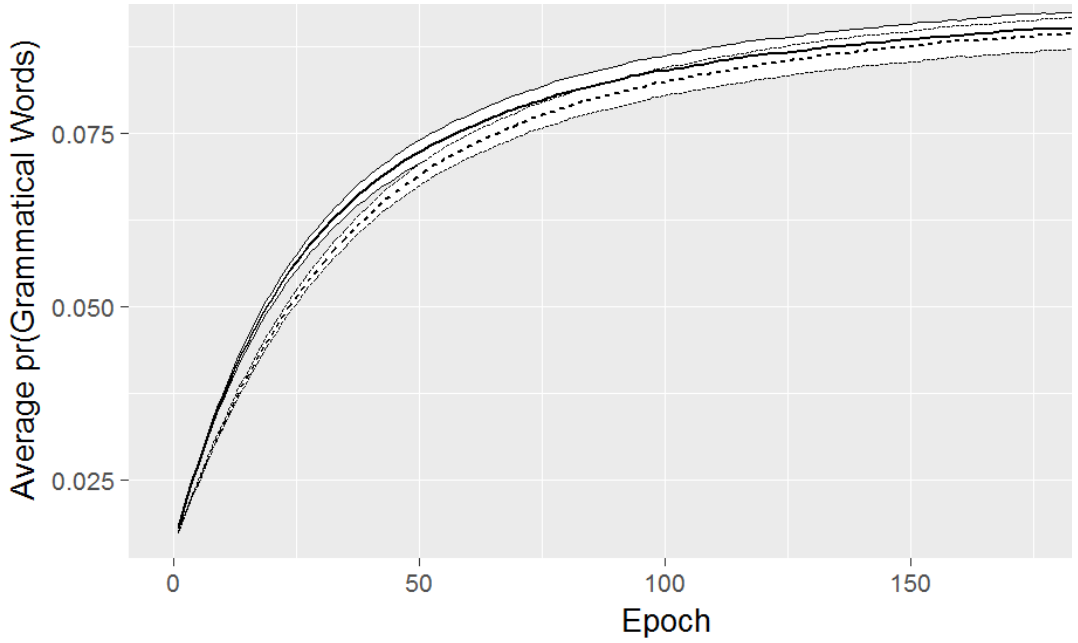


With PFA



Intradimensional Bias (Moreton 2012)

Without PFA



With PFA