Deriving frequency effects from biases in learning

Maggie Baird*

Abstract. This paper presents a phonological learner that derives frequency effects – the propensity of more frequent items undergo deletion and reduction processes at higher rates. The model is a bidirectional Maximum Entropy grammar which has two distinct learning steps, one mapping from UR to SR, and another mapping back from SR to UR using Bayesian inference. The model is tested on the case of t/d deletion in English and correctly derives the frequency-based pattern of deletion without access to surface patterns.

Keywords. MaxEnt; Bayesian inference; bidirectional grammar; phonological variation; frequency effect; probabilistic reduction

1. Introduction. Certain phonological processes are not categorical, but apply variably. In the study of variation, one important question is understanding the factors that influence the rate of application of a phonological process. This paper focuses on the effect of lexical frequency. Across languages, higher frequency words undergo higher rates of reduction and deletion processes (Hooper 1976; Bybee 2001, 2002; Bell et al. 2009, a.o.).

One explanation for why reduction processes target higher frequency items at higher rates comes from a listener-oriented perspective. In this view, there is a tension between an articulatory pressure to reduce and an "acceptable acoustic form" (Seyfarth 2014). Speakers have to balance producing a looser articulation against comprehension. Words are more likely to be understood when they are frequent or predictable, and therefore listeners are more likely to understand an incorrect or reduced form of more frequent words (Seyfarth 2014).

Along this line, I argue that frequency effects in language arise because of comprehension. Over time, speakers learn that their listeners make more errors in comprehension on reduced low-frequency items than high-frequency items. The speaker adjusts their grammar accordingly, and frequency effects are explained by errors in the comprehension process.

I present a bidirectional phonological grammar (Boersma & Hamann 2008) with learning steps based on both production and comprehension. A bidirectional grammar is a grammar which uses the same constraints and ranking in production and perception. This listener-oriented model generates fewer errors for comprehension for reduced frequent words than infrequent words. This systematic bias in the learning process derives frequency effects from unbiased data, where there are not frequency effects present. The model demonstrates how frequency effects might emerge over time, from a learning bias favoring comprehension.

The model presented here is a Maximum Entropy Grammar with lexically indexed constraints. The learning update employs Stochastic Gradient Descent and the model has two update steps: one based in production and one based in comprehension.

In section 2, I walk through the components of the model and the learning process. In section 3, I demonstrate the model using a test case from t/d deletion in American English. Section 4 discusses some comparisons to other models and section 5 concludes.

* I would like to thank groups at UMass Sound Workshop, NECPhon, and the LSA for valuable feedback. In addition, thanks to Gaja Jarosz, Joe Pater, Brandon Prickett, Andrew Lamont, Max Nelson, Seoyoung Kim and Bethany Dickerson for comments and help throughout this project. Author: Maggie Baird, University of Massachusetts, Amherst (mbaird@umass.edu).

2. Model. In this section, I walk through the setup of the model, starting with the basic frameworks of MaxEnt and lexically indexed constraints, in 2.1 and 2.2. I then present the learning algorithm which trains the model and biases the learning towards learning frequency effects, in 2.3. Finally, I discuss how data is input into the model in 2.4.

2.1. MAXENT. The data are modeled in a Maximum Entropy (MaxEnt) Grammar (Goldwater & Johnson 2003). A MaxEnt grammar is a constraint-based grammar that maps from underlying representations (URs) to surface representations (SRs). In MaxEnt, each constraint has a numerical weight, but rather than selecting one winner, the grammar generates a probability distribution over candidate SRs.

The *harmony* of a candidate is the weighted sum of its constraint violations. Candidates with harmonies closer to 0 are more probable. The probability of each candidate is calculated by applying the softmax function to the set of harmonies. Therefore, the probability of a candidate *x* is defined as follows:

$$p(x) = \frac{e^{-(\sum_i w_i c_i(x))}}{\sum_{y \in \Omega} e^{-(\sum_i w_i c_i(y))}} \tag{1}$$

where w_i is the weight of the *i*th constraint, c_i is the number of times the candidate *x* violates the *i*th constraint and Ω is the set of candidate SRs which share the same UR as *x*.

2.2. LEXICALLY INDEXED CONSTRAINTS. Lexically indexed constraints (see Pater 2010 for an overview) have been used to model a variety of phonological phenomena. In this theory, constraints can have a general version which is applicable to all lexical items or morphemes in a language, and other versions which are indexed to specific lexical items or morphemes. These constraints are only violated by the specific items, and not other morphemes. Lexically indexed constraints are ranked or weighted in addition to the general constraints. These constraints were first proposed to account for exceptionality, and have proven very successful in that domain.

My model uses lexically indexed markedness constraints. Each word can have a different rate of application of a process depending on the weighting of its lexically indexed constraint.

2.3. LEARNING ALGORITHM. This model is trained online. The error-driven learning update rule used is stochastic gradient descent. At each production learning step, the learner picks a word *x* from the set of URs. It samples this word proportionally to its frequency - that is, more frequent words will be selected as the learning datum more often.

The learner selects an observed SR from the input distribution. The learner then selects an expected SR based on the probability distribution created by the current weights of the grammar. If the observed and expected SR do not match, the weights are updated using the following update rule:

$$w_i = w_i - \lambda(O[c_i(x)] - E[c_i(x)]) \tag{2}$$

where the weight of a constraint w_i is decreased by the learning rate λ times the difference in the violations between the observed SR O and the expected SR E.

Above, I have described the standard MaxEnt learning step using gradient descent. The novel aspect of the current model is the addition of another learning step based on comprehension. I propose that frequency effects are derived from the comprehension process. Frequent

items license more reduction and deletion because they are easier to recover for the listener. This is implemented in the model through a second learning step modeling the comprehension process.

After the production-driven update, the learner takes the expected SR from the production learning step and attempts to map it back to a UR. A schematic of the learning process is presented below:

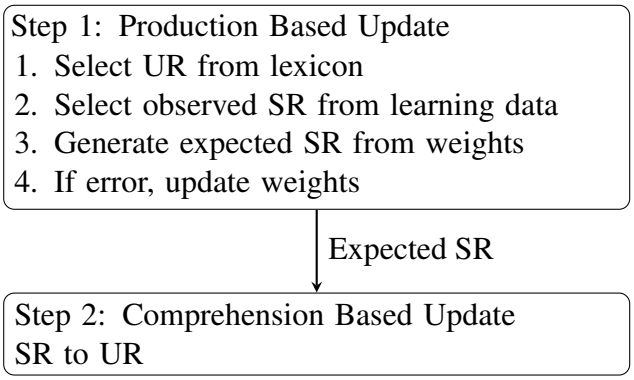

Figure 1. Step-through of the learning process

At this stage in learning, the model sees an SR, the expected output. Translating this into more realistic terms, the model is a listener who has just heard a word. It is now attempting to understand that word, by mapping it back to a lexical item, a UR.

This comprehension process is an instance of Bayesian inference. In the model, the learner is choosing a UR based on an SR, so it has to figure out the probability of a UR given that SR. The learner is therefore solving for $p(UR|SR)$.

$$p(UR|SR) = \frac{p(SR|UR)p(UR)}{p(SR)} \tag{3}$$

In the above equation, we can define the terms as follows:

$p(SR|UR)$ = The current probability assigned to the SR in the tableau of the UR given the current weights.
$p(UR)$ = The prior probability of selecting a UR from the lexicon, estimated as the relative frequency of the UR.
$p(SR)$ = The sum of the probabilities of all the possible ways of generating the current SR, the form given to the listener.

The first step is to determine which URs in its lexicon could have generated that SR. Once the model has those URs, it calculates the likelihood of each UR given the SR, $p(UR|SR)$.

In addition to considering the URs which could have generated the SR, the model/listener could posit that this utterance was an instance of a new word which they do not have in their lexicon. This will constitute an error in comprehension. To translate to a real-world case, this would be an example of a listener hearing a reduced/altered form of a word that they are unable to parse, thus positing it must be a word they haven't heard before.

When the learner posits a new word, its frequency is a pseudo-count in the lexicon. It is instantiated into the model as a parameter N, representing the pseudo-count of a new word. In the model, we need $p(UR)$ for a new word, which is calculated in the following way:

$$p(\text{UR}_{\text{New Word}}) = \frac{\text{N}}{\sum_i^n \text{freq}(\text{word}_i) + \text{N}} \tag{4}$$

Selecting the UR /New Word/ in the comprehension step represents positing the heard SR as a new lexical item with a fully faithful UR, with the relative frequency $p(UR)$ shown above. The other important measure for the formula is $p(SR|UR)$. Under the assumptions of this model, $p(SR|UR)$ will always be 1 when considering UR as New Word.

The likelihood of choosing to attribute the SR to a new word rather than the correct UR is inversely proportional to $p(UR_{correct}|SR)$, which is in turn dependent on the frequency of the correct UR.

The model creates a probability distribution over possible URs that could have generated its SR data point. It then samples from that distribution. If the selected UR is correct - that is, it is the same UR that was selected in the initial production learning step - no update occurs. If the incorrect UR is selected - either the wrong lexical UR or New Word, the weights are updated again, but the observed SR is set to the faithful form rather than the observed SR from the input data.

The reasoning behind this update is that the "listener" was not able to map back to the correct UR and thus did not properly understand the signal. In order to accommodate the "listener", the "speaker" will promote faithfulness for this word, as the faithful pronunciation will always be more easily reconstructed. Because frequent items are more likely and generate less errors, faithfulness will be promoted more, overall, for less frequent items. As learning progresses, the production step will push all lexical items to the same rate of reduction, but the comprehension update will push lexical items away from each other on the basis of frequency. The learner stabilizes on a state with different rates of application of a phonological process correlated with frequency.

Now, the step through of one iteration of the model looks as follows:

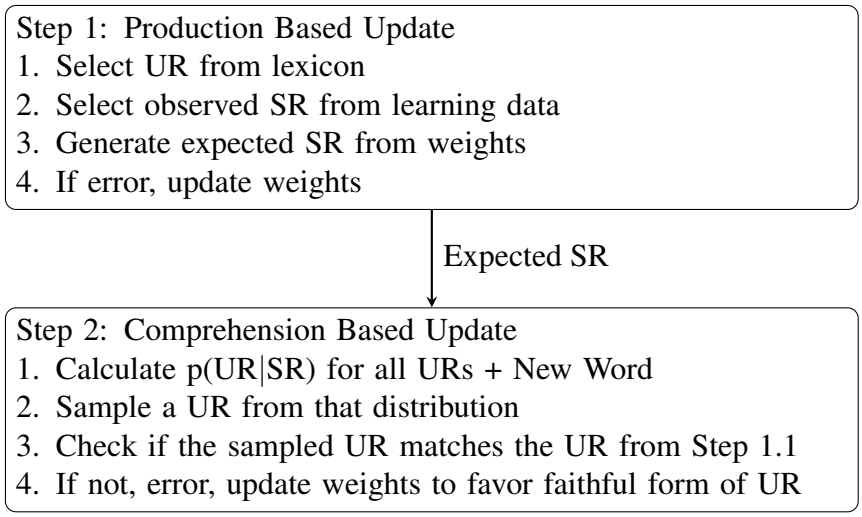

Figure 2. Complete step-through of the learning process

For each iteration of the model, there are two potential weight updates that can occur rather than one in a traditional MaxEnt model. The weights may be updated during the production step, the comprehension step, or both, depending on the errors that are generated. These updates are not necessarily in the same direction - the production based update will favor whatever the observed SR is from the input data, but the comprehension based update will favor the faithful form.

New Word is not "remembered" from learning step to learning step. In a real-world listening scenario, a listener would probably enter a new word into their lexicon, but in the model, the New Word is a tool in the learning process. The lexicon does not change over time in the learning process. It is simply comprised of the URs and corresponding SRs that are input by the analyst.

2.4. INPUT DATA. The purpose of this model is not to simulate learning constraints based on observed patterns in a synchronic sense. The model is intended to derive patterns from frequency. With lexically indexed constraints, it is a trivial learning problem to input words with rates of application of a process and run stochastic gradient descent. The learner will easily be able to represent the patterns because there is a constraint for each word, and the learner can learn the weights to capture individual words' patterns. For example, we can look at this toy example from a MaxEnt model with lexically indexed constraints trained with standard stochastic gradient descent. I trained two different languages, one with data that mimics real-life patterns (higher frequency words having a higher rate of a phonological process) and one with arbitrary data. The model, equipped with indexed constraints, is able to learn each of the patterns equally well.

Word	Frequency	Observed Rate of Reduction	Predicted Rate of Reduction
A	10000	90%	90.00%
B	1000	80%	80.00%
C	100	70%	70.00%
D	10	60%	60.00%

Table 1. Results from a MaxEnt learner on naturalistic schematic data

Word	Frequency	Observed Rate of Reduction	Predicted Rate of Reduction
A	10000	10%	10.00%
B	1000	90%	90.00%
C	100	20%	20.00%
D	10	80%	80.00%

Table 2. Results from a MaxEnt learner on unnaturalistic schematic data

These illustrative examples show that a model with lexically indexed constraints given an input distribution of observed data will be able to learn that pattern. This learning setup is not suitable for the aims of the current paper. The model presented has a goal of deriving the frequency effect through the learning process. The model should be able to take an input distribu-

tion and through the biases in the learning process, the frequency effect will emerge, even if it were not present in the input data.

In an effort to capture the emergence of frequency effects over time, the observed rate of reduction/deletion in this model is input as 100% for all words. The success of the model will be to push the words away from a deletion rate of 100%[1] and show how lower frequency words will be reduced/deleted less over the learning process. I show that the model has a learning bias towards correlating lexical frequency and rate of reduction/deletion even when that pattern is not present in the data. As such, the input data to the model look as follows:

Word	Frequency	Observed Rate of Reduction
A	10000	100%
B	1000	100%
C	100	100%
D	10	100%

Table 3. Input data setup

The words and frequencies are taken from a corpus, the specifics of which are discussed for the test case presented in this paper. The words and frequencies in Table 3 are illustrative. In the input to the model, the reduction/deletion process applies equally to every word in the lexicon. There is not a relationship between frequency and rate of application of the phonological process. The aim of the learning process is to take the distribution and show how the output does have the correct relationship between frequency and rate of application. That relationship is being derived in the learning rather than coming from the input data. This model is showing where frequency effects arise from in languages from pressures in the comprehension process - it is not an example of a learner mimicking the language acquisition process from realistic data.

3. Test Case: English t/d deletion.

3.1. DATA. To demonstrate the model, the model was tested on data from t/d deletion in English. In English, word-final t/d variably deletes from consonant clusters. For example, the word *went* may be pronounced as either [wɛnt] or [wɛn].There is a frequency effect at work, with more frequent words undergoing deletion more often than less frequent words - for example, *just* deletes more than *jest* (Patrick 1992; Bybee 2000; Phillips 2006; Coetzee 2009; Coetzee & Pater 2011; Coetzee & Kawahara 2013).

Data are taken from Coetzee & Kawahara (2013) (henceforth C+K). Their source is the Buckeye corpus (Pitt et al. 2007), a corpus of conversational English spoken in Columbus, Ohio. The corpus consists of 40 speakers and over 300,000 tokens. The data in the Buckeye corpus are phonetically transcribed. Every utterance has both a phonemic and a phonetic transcription. C+K took all words which orthographically ended in Cd and Ct and manually removed past tense words, tokens ending in rd/rt/lt/ld (see C+K for justification), and words

[1] The actual rate of 100% is relatively arbitrary. I chose it to represent the articulatory pressure to delete - all words have the same inherent force to delete/reduce. I have also run the model with the observed rate being the average rate of deletion across the language. In the test case presented in this paper, the 100% does better. The important component is that the rate of deletion is the same for all the input data.

which orthographically ended in clusters, but not phonemically (such as 'would'). After this filtering process, they were left with 16,460 tokens and 459 types.

The Buckeye corpus is a relatively small corpus, and thus it is difficult to estimate rates of deletion for low frequency words, as they have few tokens in the corpus. In order to deal with this problem, C+K binned the data to get a more reliable estimate of how deletion rates change as a function of frequency. They obtained the CELEX (Baayen et al. 1995) frequency of each word and took the logarithm (base 10). They created bins at intervals of 0.1 on the logarithmic scale, and combined bins with fewer than 50 tokens with adjacent bins, for a final grouping of 23 bins. These bins ranged from (0 to 2.0) to (5.7 to 5.8) on the log frequency scale. C+K calculated the rates of deletion for each of these bins in three phonological contexts (pre-consonantal, pre-vocalic, and pre-pausal), which they show have different rates of deletion. There is a positive correlation between frequency and rate of deletion in each of the three contexts.

In this case study, I model the same t/d data that C+K did in the Pre-C condition. That is, instead of inputting individual words to my model, the lexicon is comprised of the 23 bins of words they created. The lexically indexed constraints are indexed to bins rather than individual words. This input representation is not a core part of the model, but rather a necessity for the data available. In theory, the URs and SRs that the model is working with are phonological contentful. This specific dataset requires us to abstract away from that and model over bins rather than individual words. In this case, URs are not comprised of phones, bur rather look like this: /bin X/, and SRs look like this: [bin X], [binX-deleted]. The possible SRs for a given lexical item go beyond just a faithful and deletion candidate in theory, but these are the only two candidates we are concerned with in the modeling, and are therefore the only candidates included for the simulations. The lexicon, therefore, only contains 23 items. Each of these items has a corresponding frequency.

Although the bins are defined by log frequency, the frequency of each bin in the input data was raw frequency taken from CELEX. As the bins may contain more than one word with different frequencies, I chose the highest corresponding frequency to represent that bin. For example, the '2.2' bin contains words ranging from CELEX frequency 104 'disagreement' to 148 'easiest'. In the input lexicon file, the UR /bin 2.2/ has a corresponding frequency of 148.

The model uses the same constraints as C+K, except that the markedness constraints are indexed. There is one general markedness constraint (*CT$]_{Word}$), one general faithfulness constraint (MAX), and 23 indexed markedness constraints (*CT$]_{Word}$ (bin X)).

The parameter for New Word was set to 5000, after testing various other values, so the p(new word) was calculated as follows:

$$\frac{5000}{\sum_{i=1}^{n} freq(bin_n) + 5000} = 0.008 \tag{5}$$

3.2. RESULTS. In the data, the most frequent bin has a frequency of 514946, and the smallest bin has a frequency of 91. In online learning, each learning datum is chosen from the underlying frequency distribution. Because of the wide range of frequencies, and the fact that the highest bin accounts for nearly 84% of the data, the learning process is relatively long. The results reported below required 2,000,000 iterations of the learner. The weights were all initialized at 1, and the learning rate was set to 0.02. The results of one run of the model are pre-

sented in Figure 1. The y axis shows the rate of deletion and the x axis shows the bins, by log frequency. The blue line is the model's predictions and the black dotted line is the actual data from Buckeye.

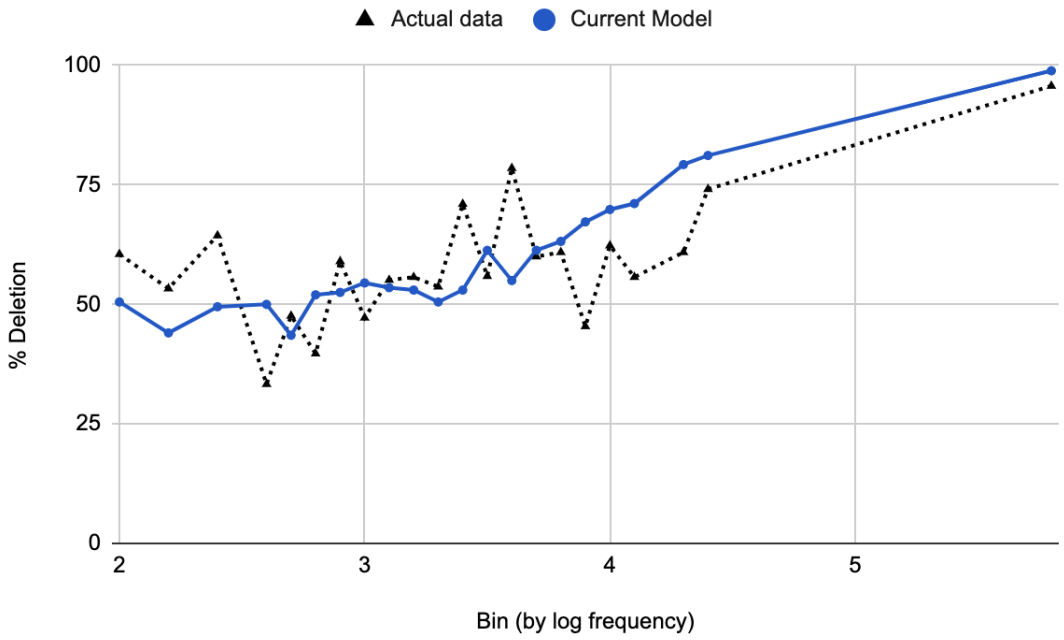

Figure 3. Predicted Percentage Deletion from a model run vs. corpus data

The most important benchmark is that the model correctly derives a positive relationship between frequency and rate of deletion, and it clearly does. Recall that in the input data, all bins have the same rate of deletion of 100%. As we can see from Figure 1, the observed data does not follow a strictly increasing relationship between frequency and rate of deletion. Individual points deviate strongly from the pattern, such as the lowest frequency bin showing a much higher rate than some of the higher frequency data. The goal of the model is not to perfectly match the Buckeye data, but rather to model the trend we observe based on frequency. The Buckeye data is a decent approximation of this, but it is a small corpus, and the frequency bins contain small numbers of tokens in some cases. It is perhaps the best estimate of actual deletion rates that we have, but it is not perfect.

There is not one perfect quantitative metric on which to evaluate this model, however, we can report some quantitative measures. The Pearson's correlation coefficient between the corpus data and the model's output for this run is 0.602. Another measure I used to evaluate success was a metric from the C+K paper, which was Mean Squared Error improvement over baseline. The baseline there was a Harmonic Grammar without scales, which learned one rate of deletion for all the data, which was 79.4%. The MSE improvement from their modeling was 75.83%. The average MSE improvement over the same baseline for my model across 20 runs was 76.96%. This quantitative metric is not perfect as the learning model here and the account in C+K have very different approaches (see section 4 for more details) Still, it shows that the model does about as well as other analyses at capturing surface patterns, even though it is not the primary goal. The crucial success is deriving the relationship between frequency

and reduction.

The predictions of my model and C+K's are presented below in Figure 4, with the baseline in black, the corpus data in blue, my model in green and C+K's in yellow.

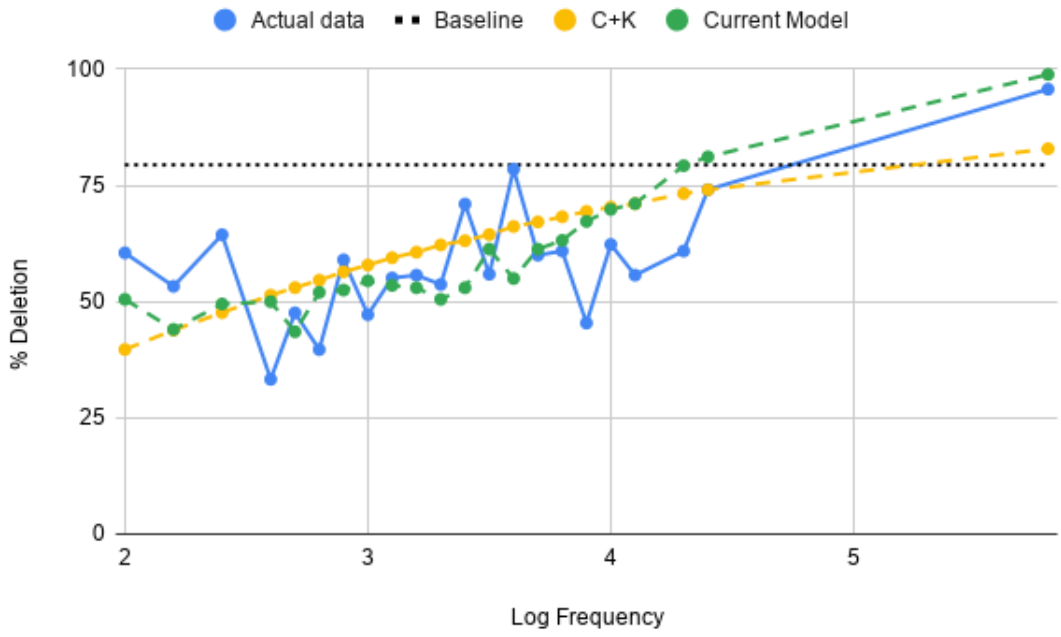

Figure 4. A comparison of my model and C+K's results

With an online learner with so many constraints, there will be variation in the output, and the learner does not always behave in the same way. In a successful model, the relationship between frequency and rate of deletion should always emerge. In Figure 5, I show the behavior of the model over 20 runs. For each bin, the highest predicted rate and the lowest predicted rate are plotted from the 20 runs, against the observed data. As the minimum, maximum and median predicted rate are calculated for each bin, the lines in the graph do not represent singular runs of the model - the values can be from different runs for each given bin.

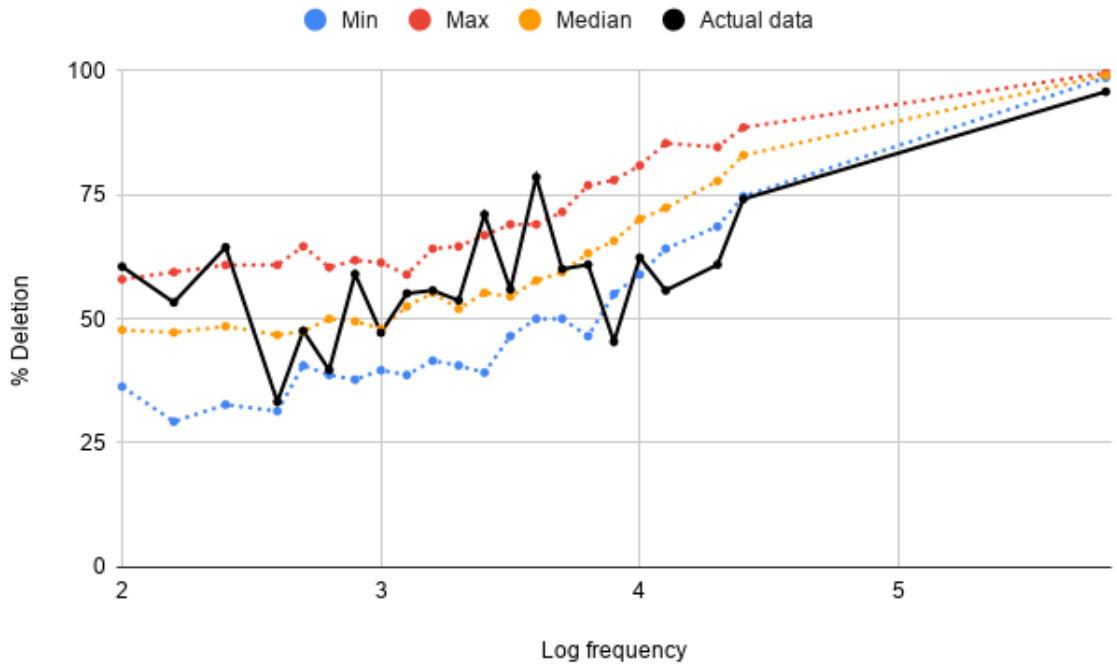

Figure 5. Minimum, maximum and median predicted rates from 20 runs of the model

4. Discussion. In 3.2, I report evaluative metrics from the C+K paper. In their paper, they propose a linking function from frequency to constraint scales, which change the weight of constraints based on the frequency. This linking function is defined such that lower frequency words will always have higher faithfulness than higher frequency words. The relationship between frequency and faithfulness is built into their model. The constraints in my model have no such built-in preference. In 2.4, I showed how a MaxEnt model with lexically indexed constraints can learn a pattern with a relationship between frequency and application of a phonological process. The same learner can just as easily learn an arbitrary pattern.

My model has the capacity to learn arbitrary relationships between frequency and phonological processes. It is the comprehension process that biases the learning towards a positive relationship between frequency and rate of application. This pattern emerges even when the input data does not have the relationship and when the constraints have the power to learn entirely different patterns. The model derives the pattern without having it built in.

This distinction makes the goals of my and C+K's paper very different. Their paper attempts to capture surface patterns, while mine shows how the patterns on the surface arise over time from unbiased data. The equivalent performance our models have on capturing surface patterns is therefore a huge success. My model does as well as theirs at capturing the surface data, without having seen any of the surface data.

There are still many limitations of this learner which need to be addressed in future work. Notably, not all phonological processes show the same frequency effects - these effects appear to be much more common with reduction and deletion processes. Although this is what I have modeled, there is no principled component of the model which targets these processes specifically. In order to expand the model further, I would also want to consider more SRs than sim-

ply the faithful and deleted candidate. Along this line, New Word is always considered to be a faithful version of the SR, but this is of course a simplifying assumption. Furthermore, the frequencies of different lexical items are given to the learner, but it would be interesting to investigate how the learner would behave if it were learning frequencies over time as it "heard" the different lexical items according to their natural distribution.

As this is a project intended to model the emergence of frequency effects over time and the goal is to show how languages develop over time, a natural direction for this model is to look at iterated learning and learning over generations (Kirby & Hurford 2002; Wedel 2011; Moreton & Pater 2012). Currently, the model starts from having the same observed rate of deletion for every lexical item. With iterated learning, I could start at the same point, but stop learning after a certain number of steps and use the output of that model as the input distribution to the next and simulate a language evolving over time, from generation to generation.

5. Conclusions. The model presented in this paper shows promise as an explanatory model of frequency effects based on comprehensibility. Listener-orientedness has long existed as an explanation for frequency effects in phonological processes. This paper implements this hypothesis in a computational learning model which derives frequency effects in a bidirectional version of a standard MaxEnt learner.

In this paper, I have presented a new model which is able to explain frequency effects in phonological processes based in reduction and deletion. The model is able to capture real language data as well as other proposed models, which was shown with a case of t/d deletion in English data. It also has explanatory power in showing how the patterns emerge from biases in learning, using a language-general explanation rooted in comprehension.

This paper presents both a way of capturing surface linguistic patterns, but primarily serves as a way of explaining and deriving why those effects arise. I argue that frequent words are more reduced because they are easier to recover for the listener. The modeling work supports this explanation, and shows how a neutral lexicon is biased towards frequency-based reduction patterns through the comprehension-based learner.

References

Baayen, R. Harald, Richard Piepenbrock & L. Gulikers. 1995. CELEX2 LDC96L14. Web download. Philadelphia: Linguistic Data Consortium. https://doi.org/10.35111/gs6s-gm48.

Bell, Alan, Jason M. Brenier, Michelle Gregory, Cynthia Girand & Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 92–111. https://doi.org/10.1016/j.jml.2008.06.003.

Boersma, Paul & Silke Hamann. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25. 217–270. https://doi.org/10.1017/S0952675708001474.

Bybee, Joan. 2000. The phonology of the lexicon: evidence from lexical diffusion. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 65–85. Stanford: CSLI Publications.

Bybee, Joan. 2001. Frequency effects on French liasison. In Joan Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 337–369. Amsterdam: John Benjamins.

Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14. 261–290.

Coetzee, Andries. 2009. An integrated grammatical/non-grammatical model of phonological

variation. In Young-Se Kang, John-Yurl Yoon, Hyunkyung Yoo, Sze-Wing Tang, Yong-Soon Kang, Youngjun Jang, Chul Kim, Hyoung-Ae Kim & Hye-Kyung Kang (eds.), *Current issues in linguistic interfaces* (vol. 2), 267–294. Seoul: Hankookmunkwasa.

Coetzee, Andries & Shigeto Kawahara. 2013. Phonological variation and lexical frequency. *Natural Language and Linguistic Theory* 31. 47–89. http://www.jstor.org/stable/42629730.

Coetzee, Andries & Joe Pater. 2011. The place of variation in phonological theory. In Alan Yu John Goldsmith, Jason Riggle (ed.), *2nd Edition of the handbook of phonological theory*, 401–431. Blackwell.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Variation within Optimality Theory: Proceedings of the Workshop on Variation within Optimality Theory*, 111–120. Stockhom: Stockholm University.

Hooper, Joan Bybee. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In William M. Christie (ed.), *Current progress in historical linguistics*, 96–105. Amsterdam: North Holland.

Kirby, Simon & James R. Hurford. 2002. The emergence of linguistic structure: An overview of the iterated learning model. In Angelo Cangelosi & Domenico Parisi (eds.), *Simulating the evolution of language*, 121–147. Dordrecht: Springer.

Moreton, Elliot & Joe Pater. 2012. Structurally biased phonology: Complexity in learning and typology. *Journal of the English and Foreign Languages University (Hyderabad)* 3(2). 1–44.

Pater, Joe. 2010. Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In Steve Parker (ed.), *Phonological argumentation: Essays on evidence and motivation*, 123–154. London: Equinox.

Patrick, Peter L. 1992. Creoles at the intersection of variable processes: -t, d deletion and past-marking in the Jamaican mesolect. *Language Variation and Change* 3(2). 171–189. https://doi.org/10.1017/S095439450000051X.

Phillips, Betty S. 2006. *Word frequency and lexical diffusion*. New York: Palgrave Macmillan.

Pitt, M.A., L. Dilley, K. Johnson, S. Kielsling, W. Raymond, E. Hume & E. Fosler-Lussier. 2007. Buckeye corpus of conversational speech (2nd release). Columbus, OH: Ohio State University (Distributor). https://buckeyecorpus.osu.edu/php/publications.php.

Seyfarth, Scott. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133(1). 140–155. https://doi.org/10.1016/j.cognition.2014.06.013.

Wedel, Andrew. 2011. Self-orgnization in phonology. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell companion to phonology*, 130–147. Hoboken, NJ: John Wiley and Sons.