Phonotactic Learning with Structure, not Statistics

Logan Swanson, Jeffrey Heinz, Jonathan Rawski

Abstract

We provide empirical evidence against Wilson & Gallagher's 2018 claim that statistics is necessary for phonotactic learning. We implement BUFIA, a feature-based non-statistical learner and, using the same data and case study as Wilson & Gallagher, show that this non-statistical learner is equally successful at learning phonotactics as the Maximum Entropy-based learner they use. This counters their conclusion that non-statistical phonotactic learning is impossible, while supporting their advocacy for feature-based representations in phonotactic learning.

Keywords: computational phonology, phonotactic learning, abductive inference

In a phonotactic learning task, a learner must tackle two difficult problems simultaneously: first they must be able to *identify* constraints that are consistent with the data they receive. Second, they must be able to *select* which of these constraints to actually incorporate into their grammar. For feature-based representations, where many different constraints can account for the same series of segments, this second task is particularly important.

Wilson and Gallagher (2018) (henceforth W&G) argue that both statistical methods and feature-based representations are essential ingredients for phonotactic learning. They provide an experimental study of accidental gaps in Bolivian Quechua as evidence, using the MaxEnt phonotactic learner introduced by Hayes and Wilson (2008), which

accomplishes constraint selection by incorporating the statistical principle of maximum entropy and its associated techniques into the model, alongside other heuristics and parameters. Their study considers segment-based models (both statistical and non-statisical) as a point of comparison and shows that the MaxEnt learner outperforms both.

We investigate W&G's claim that statistical methods are necessary for phonotactic learning using the Bottom-Up Factor Inference Algorithm (Chandlee *et al.*, 2019) (BUFIA), a non-statistical, deterministic algorithm which leverages the rich internal structure of a feature-based phonotactic constraint space to find the most general constraints consistent with the data. Using the same data, feature-based representations, and training setup as W&G, we find that BUFIA performs competitively with the MaxEnt model used by W&G, in some ways actually outperforming it.

Specifically, we run two sets of experiments. The first set replicates the studies in W&G, adding BUFIA to them. The second set of experiments addresses potential confounds in the test sets of the first set of experiments. In both sets of experiments, BUFIA's performance compares favorably to W&G's MaxEnt model. Based on these results, we agree with W&G that feature-based representations are important for phonotactic learning but dispute the necessity of statistical techniques. Our results show that statistical methods are not the only way to achieve success in phonotactic learning, and that *structure* can play a vital role.

The remainder of this reply is organized as follows. Section 1 discusses the role of both statistics and representation in learning, lays out the paradigm for comparing models presented by W&G, and explains how BUFIA makes this paradigm complete. Section 2 briefly recaps how BUFIA operates and discusses some of its technical properties. Section 3 turns to the Quechua case study and outlines the two sets of experiments that were conducted to assess BUFIA's performance. Section 4 discusses the experimental

findings and how they bear on our understanding of phonotactic learning. Section 5 offers ideas for future directions and concludes.

1. Ingredients for Learning

W&G investigated the impact of two main factors on learning: representational choice and statistics. They compared three learning models: a feature-based MaxEnt model, a segment-based MaxEnt model, and a segmental non-statistical model. Based on the performance of these three models at learning phonotactic patterns in Bolivian Quechua, they conclude that both *featural representations* and *statistical methods* are essential for learning.

As shown in Table 1, there is a fourth logically possible type of model which would complete this paradigm–one which employs featural representations but no statistics. W&G consider this fourth possibility, but conclude that the exponential number of possible featural representations corresponding to any given segment sequence renders such a model impossible (p. 617):

> *"Lacking a method for deciding which representations are relevant for deciding well-formedness – precisely the role played by statistics in MaxEnt-Ftr – learning... is doomed."*
>
> –Wilson and Gallagher (2018)

However, more recent work has since shown that just such a model does in fact exist. The Bottom-Up Factor Inference Algorithm (Chandlee *et al.*, 2019) is deterministic, non-statistical, and can operate over featural representations. This algorithm relies on the structure of the constraint space itself to make learning possible.

2. BUFIA: a non-statistical feature-based approach

2.1. Representational Considerations

BUFIA expresses constraints as *relational structures*, a well-known area of mathematical logic and model theory for discussing structures and containment. Provided an appropriate representational scheme, distinct linguistic units, such as distinct words, correspond to distinct relational structures. Relational structures naturally provide a notion of containment: a relational structure A contains another structure B provided there is a way to map elements of B into A such that the properties and relations that hold among the elements in B hold among the corresponding elements in A. (See Chandlee *et al.* (2019) and Rogers and Lambert (2019) for mathematical details.)

For example, consider the relational structure that consists of two ordered events, where the first has the property of being a front vowel, and the second has the property of being a back vowel, for which the expression [−back][+back] could be used. This structure is contained in the structure for the string [iu]: map the first element to [i], the second element to [u], and the properties and ordering relations that hold for the elements in the structure for [−back][+back] hold for the corresponding elements in the structure for [iu].

Following earlier work, we refer to these relational structures as "factors." If structure B is contained in structure A, we also say that B is a *subfactor* of A, and A is a *superfactor* of B.[1] Relational structures and the containment relation naturally give rise to the notion of markedness constraints. If structure B is marked, then each of its superfactors contain a marked structure.

Recent phonological work has examined segmental structures, tone (Jardine, 2017), stress (Lambert, to appear) , syllable structure (Strother-Garcia, 2018), and prosodic and morphological structur (Dolatian, 2020) through the lens of relational structures. Stating

BUFIA in terms of factors means that it can be applied to all of these representations, and anything else formulated over relational structures (Section 4.3 further discusses BUFIA's generality).

2.2. Leveraging Structure for Learning

With constraints represented as sequences of bundles of feature values, the exponential combination of features yields a search space that grows exponentially as the number of features increases. However, this space is not a random grab-bag, but rather has a rich internal structure, forming a partially ordered hierarchy. As a working example, consider the simple toy phonological system given in Table 2, adapted from Rawski (2021). This system contains just four vowels, parameterized by two features with binary $\pm$ values.

The possible constraints over this system (represented by sequences of bundles of features) correspond to banned *subfactors* of possible words. Because of this, the space of possible constraints can be organized into a hierarchy by the *containment* relation with superfactors dominating their subfactors. Figure 1 gives a fragment of this partially ordered space of possible constraints. Under this containment relation, the space of possible constraints grows upwards like an inverted pyramid. At the bottom, there is an "empty" factor which matches (i.e., is a subfactor of) all segments: [ ]. Each consecutive "layer" moving upwards in the hierarchy contains structures which can be generated from the layer below by adding a single segment (i.e., an empty feature bundle) or a single feature. Notice that in this hierarchy, *licitness* proceeds *downwards* while *illicitness* proceeds *upwards*: the licitness of a factor implies the licitness of all its subfactors, while the illicitness of a factor implies the illicitness of all of its superfactors. So, if the factor [-back][+back] is determined to be illicit, it follows that its superfactor [-back][+back, +high] must also be illicit. Similarly, if the factor [-high][+high] is licit, then it follows that all of its subfactors, such as [-high] and [+high], must also be licit.[2]

To learn phonotactic constraints, BUFIA conducts a breadth-first traversal of this hierarchy, starting from the bottom and using the containment relationships between possible constraints to prune the search space as it goes. It takes as input a sample of positive data, and outputs a grammar of constraints which specify banned structures. A basic outline of the algorithm's steps are as follows:

1. Begin at the bottom of the space of possible constraints with an empty grammar, and proceed upwards layer by layer
2. For each structure that is encountered, check to see if that structure is present in the input data.
    - If it is, it is licit, so continue the traversal
    - If it is not, it is illicit, so add it as a constraint to the grammar, and prune *all superfactors* of that structure out of the remaining search space
3. Stop at a particular cutoff condition. Typically, this will be a limit on the "size" (in terms of sequence length and number of features per bundle) of constraint that can be considered, but it could also be some other stopping condition like total number of constraints.

For this toy phonotactic example, suppose BUFIA is presented with an input sample consisting of all strings of length two where the segments match in backness (enumerated in Table 3). The algorithm will begin at the bottom of the constraint space, at the empty feature bundle, which is the root. Since this is a subfactor of all segments, it is trivially contained in the input, and BUFIA will proceed to the next "layer" of the partially ordered space. This next layer contains just those factors which can be generated from the root by either adding one feature or appending one segment slot (empty bundle): [+high], [-high], [+back], [-back], and [][] (not depicted in Figure 1). Each of these factors is again present

in the data, so the algorithm will proceed upwards to the next layer, which includes all the factors obtained by either adding one feature or appending one segment slot to any of the factors in the current layer.

This process continues until BUFIA encounters a factor which is *not* present in its input sample. In this case, while searching the third layer shown in Figure 1, it will find that the factor [-back][+back] is absent from the data. When this happens, the algorithm will first add the missing factor as a constraint to its grammar, and then it will prune the search space so that no superfactors of [-back][+back] are later considered. Once the stopping condition is met (typically, once all constraints up to a certain size have been considered), the algorithm will halt and return its grammar, which will contain all and only those constraints it has found. In the case of our toy example, this will be just three constraints: *[-back][+back], *[+back][-back] (these ensure that adjacent segments match in backness), and *[][][] (this forbids any string longer than two segments).

The step of pruning away superfactors once a constraint has been found is the core of BUFIA's real-world tractability. The same pattern of exponential growth which makes the overall constraint space so large also means that the parts of the space which can be pruned away are commensurately large. This means that sparse, gappy input data (a core property of human languages across domains) is highly advantageous for BUFIA, enabling it to find general constraints which substantially reduce the number of structures it must consider. Figure 2 depicts BUFIA's traversal more abstractly, and illustrates this intuition.

The MaxEnt family of learners also contain particular choices for filtering out constraints. This enables a precise comparison of the role of statistical inference: is it necessary, or are heuristics alone sufficient? The framework laid out by Wilson & Gallagher in their case study of Bolivian Quechua provides an opportunity to evaluate the performance of BUFIA and to complete the four-way paradigm they introduce, allowing for a more complete exploration of the role of representation and statistics in phonotactic

learning.

2.3. Further Learning Guarantees of BUFIA

BUFIA (like MaxEnt) is a batch learner which learns exclusively from positive data, so a typical input will be a list of licit words or forms from the target language. Chandlee *et al.* (2019) prove that BUFIA can learn constraints defining classes of local and long-distance phonotactics, and prove some relevant properties regarding the algorithm's behavior. Firstly, given a finite positive data sample D, the grammar G returned by BUFIA is guaranteed to be consistent with the data. That is, D will be a subset of the language generated by G, $L(G)$ (here, "language" is used in the formal sense: $L(G)$ is the set of all strings which satisfy all constraints in G). Additionally, $L(G)$ will be the *smallest* language in the relevant class (as defined by the types of constraints which may be considered) for which this is true. Finally, G will contain the most general factors of any other grammars G' which satisfy these first two points. This is not to say that G is uniquely the most general grammar, only that no other grammar G' where $L(G') = L(G)$ contains a factor which is more general than *any* factor in G.

These learning guarantees make BUFIA highly interpretable. Furthermore, as a deterministic algorithm, BUFIA will always return the same output when presented with the same input sample. For any constraint in a grammar generated by BUFIA, it is easy to understand how it got there: the grammars contain exactly those constraints which are as *general* as possible while still being *true* on the input data.

3. Case Study: Bolivian Quechua

3.1. Quechua Phonotactics

W&G use data from South Bolivian Quechua to conduct their experiments on the respective roles of features and statistics in learning. South Bolivian Quechua (henceforth, Quechua) is a Quechuan language spoken in Bolivia and adjacent areas of Argentina (Torero, 1964, as cited by Gallagher (2016)). Quechua roots primarily consist of two syllables, each with mandatory (simple) onsets and optional codas. In almost all cases, codas must be continuants or nasals (Gallagher, 2011). The Quechua consonant inventory includes both (phonemic) aspirates and ejectives, and there are strict restrictions governing which consonants these may co-occur with. For example, a word may not contain two aspirates or two ejectives (Gallagher, 2011). Additionally, velar and uvular consonants may not co-occur within a morpheme, and certain segmental sequences ([wo] and [wu]) are banned. Finally, Quechua displays a vowel allophony process, where the underlyingly high vowels /i/ and /u/ lower to [e] and [o] respectively in the vicinity of a uvular. These mid vowels surface always and only when there is a uvular either immediately adjacent to the vowel, or following the vowel across an intervening coda (Wilson and Gallagher, 2018).

W&G group these generalizations into surface-based constraints on four tiers which, taken together, enforce the known phonotactic generalizations active in Quechua. Table 4 summarizes the constraints and corresponding tiers, as grouped by W&G.

To represent the vowel allophony process over the dorsal tier with surface-based constraints, it is necessary to use *trigram* constraints (i.e., sequences of three feature bundles) to prohibit all configurations where a mid vowel surfaces with no uvular segment occurring on either side. W&G point out that a phonotactic learner which must consider trigram constraints (in contrast to only considering bigram constraints, which much of the

prior work has focused on) is more vulnerable to the problem of accidental gaps, since many trigrams are likely to be missing from the language purely by chance. In fact, of the 2,966 possible trigrams which are permitted according to the known constraints of Quechua, only 1,427, or just under half, are actually attested in the language.

W&G highlight an example of two trigrams containing segments with similar frequencies, [k^hek] and [eq^ho], which are both unattested in Quechua. With [k^h] and [q^h] both being positionally infrequent and [e] being the rarest surface vowel, it is perhaps unsurprising that these two trigrams would be missing. However, the sequence [k^hek] is missing for a principled reason–vowel allophony prohibits the mid vowel [e] from occurring without a licensing uvular on either side, while [eq^ho] is missing presumably purely by accident. A successful phonotactic learner, then, must be able to distinguish these two types of gaps and predict which forms are missing from the input data because there is a principled rule against them and which are missing purely by chance.

3.2. Experiment 1

To add BUFIA[3] to the paradigm given in Table 1, we conducted an experiment replicating the experimental setup used by W&G. One difference between our replication and their experiments comes from splitting the test set into different subsets for tuning hyperparameters and for evaluation. Both the MaxEnt models and BUFIA have hyperparameters that can be estimated from the data. W&G acknowledge they set hyperparameters for the MaxEnt models, but did not explain how they did so. Therefore, in our replication we divided the testing data into two random halves, one for tuning and one for evaluation (see below). We note this difference did not hurt the MaxEnt model's performance.

Following W&G, the training data for this experiment consisted of 1,104 attested roots compiled from the Laime Ajacopa dictionary (Ajacopa *et al.* 2007, as cited in Wilson and

Gallagher 2018). These forms were combined with three known suffixes (with vowel lowering applied where applicable) or left bare, resulting in a total training set of 4,416 forms. W&G randomly divided this data into five cross-validation folds, each fold consisting of 80% of the data for training, with a unique 20% held out for testing. We use the same folds in this experiment. In addition to the held-out dictionary forms, the testing data also included an exhaustive list of all possible CV(C)CV(C) sequences. These synthetic forms were divided into "licit" and "illicit" categories according to the known phonotactic generalizations given in Table 4.

For each fold, W&G trained the models on the training data and evaluated them on the testing data. As mentioned, in our experiment, we further divided W&G's testing data into two random halves, one for tuning and one for evaluation. Thus, while W&G use a two-way 80/20 train/test split, we have a three-way 80/10/10 train/tune/eval split, a standard practice for evaluating learning models with hyperparameters (James *et al.*, 2023). Due to their nondeterministic nature, each simulation for the MaxEnt models was repeated three times, and results from these simulations were averaged. The other models were run only once for each fold, since they are deterministic. Using the tuning set, the relevant hyperparameters of each model were adjusted to optimize performance. The results reported reflect performance of the model on the evaluation set with these hyperparameter values.

For both BUFIA and the MaxEnt models, the hyperparameters being tuned corresponded to the stopping condition–the point at which the learner would stop adding new constraints to the grammar. For BUFIA this could mean a number of constraints, or a certain maximum constraint complexity. For MaxEnt, this was the gain threshold required for new constraints to be added.

Table 5 compares BUFIA's performance in this experiment to that of the two MaxEnt models and the (T)SL learner explored by W&G. The results for MaxEnt-Ftr,

MaxEnt-Seg, and the (T)SL model largely align with those reported in W&G. As W&G note, all models are able to perform well on the held-out attested forms. During tuning, we found that the Maxent models which performed the best on some folds were those which had an earlier cutoff for adding constraints (i.e., a higher minimum gain required for any constraint to be added) – a 'minGain' of 200 rather than the setting of 100 used by W&G for all folds. This likely accounts for the small discrepancies between the Maxent performance reported here and those in the original W&G paper. On the synthetic testing data, BUFIA performs competitively with MaxEnt-Ftr with respect to differentiating unattested licit forms from illicit ones.

The setup of Experiment 1 was designed to highlight learning behavior on accidental gaps. However, it comes with some unintentional artifacts which could bias learner behavior. Firstly, each root in the lexicon is repeated four times in the training data, but this is not controlled for in fold construction. Because of this, many roots are present in both the training and testing sets for a given fold, meaning that models may perform artificially well on the held-out forms because of their familiarity with the roots (and suffixes) they have seen during training. Additionally, because the list of illicit forms is exhaustive (for two-syllable words), there is an uneven distribution in how many illicit forms violate each constraint. For example, a single constraint on the dorsal tier (*[+wb][-high,-low][+high], with [+wb] matching the word boundary symbol) rules out over 76,000 of the synthetic illicit forms, while other known surface constraints (for example, those enforcing syllable structure) rule out none.

Moreover, the synthetically created "licit" data is unverified by native speakers. This bakes in an assumption that the constraints described in Table 4 are the only ones active in the grammar. While this may be true, it is also possible that there are other phonotactic restrictions operating in Quechua which phonologists have not (yet) identified. From the perspective of the learning model, the held-out attested forms serve equally well to probe

behavior on accidental gaps, since these are indeed forms which the learner has never encountered which are nonetheless unambiguously permissible in the target language. To ensure these possible confounds did not introduce a bias in Experiment 1 towards any type of model, we designed an alternate training and evaluation setup and conducted a second experiment.

3.3. Experiment 2

In the second experiment, training was conducted on the same set of attested dictionary forms. However, these were divided into an alternate 5-fold split, with no roots duplicated across training and tuning or evaluation data within a given fold. For illicit data, 40 forms were randomly generated which uniquely violate each of the known constraints summarized in Table 4. Rather than rely on synthetic "licit" data, the tuning and evaluation data was limited to the held-out dictionary forms from each fold and a constraint-balanced list of illicit forms. This had the additional effect of equalizing the number of licit and illicit forms in the tuning and evaluation sets, making it possible to meaningfully report precision, recall, and F1 score, which are defined and discussed below.

Each model was trained on each fold of training data, tuned on half the held-out dictionary forms and half the curated illicit data, and evaluated on the corresponding other half of each. Once again, for the MaxEnt models, three simulations were run per fold to account for the nondeterministic nature of the learning algorithm.

Table 6 shows the results of this experiment. Reported precision, recall, and F1 score are averaged across all folds and simulations. Precision here is the percentage of total accepted forms which were part of the held-out licit data, and recall is the percentage of held-out forms which are accepted (raw percentages of accepted amounts are also included, and it is easy to see that these two numbers (in the second and fourth columns)

are identical). Broadly speaking, permissiveness will favor recall, while restrictiveness will favor precision. For example, the TSL model is able to achieve perfect precision by banning almost everything – all the illicit forms and more than half of the licit ones. The MaxEnt-Seg model does the opposite, acheiving perfect recall by *allowing* nearly everything. Because of this, F1 score (which represents the harmonic mean of these two) is a metric well-suited to capturing how well a model does at balancing restrictiveness and permissiveness. Again, BUFIA's performance is competitive with that of MaxEnt-Ftr, outperforming it in terms of overall F1 score.

4. Discussion

4.1. Structural and statistical inference

The results obtained from both Experiment 1 and Experiment 2 demonstrate that BUFIA is just as effective as MaxEnt-Ftr for phonotactic learning, at least on the Quechua dataset. The grammars generated by BUFIA are able to successfully rule out forms which violate the rules of Quechuan phonotactics, while still allowing unseen licit forms.

The success of BUFIA supports W&G's claim that feature-based representations are important for phonotactic learning. However, it undermines their claim that statistics are a *necessary* ingredient. The core insight captured by BUFIA is that the search space of phonotactic constraints is highly structured, in a way that can be leveraged for learning. In fact, this structure is referenced by the Hayes & Wilson MaxEnt learner as well. They employ a *generality* heuristic, ensuring that constraints with fewer segments and covering larger natural classes will be considered first (Hayes and Wilson, 2008).

The presence of this and other search heuristics used by the MaxEnt model makes it difficult to tell how much of its behavior is due to its statistical properties versus its non-statistical "kernel." While an in-depth qualitative comparison of the grammars

produced by MaxEnt and BUFIA is outside the scope of this paper, Rawski (2021) observes that both BUFIA and the Hayes & Wilson MaxEnt model yield identical grammars on a toy example similar to the one given in Tables 2 and 3.

These experiments show that both a purely structure-based approach (BUFIA), as well as a search combining structure with statistics (MaxEnt), are sufficient for phonotactic learning. Structure, therefore, should not be underestimated or downplayed in the development of algorithms for learning grammars. With structure alone providing a sufficient condition, the role of statistics in a combined model like MaxEnt-Ftr is left an open question to be further explored.

4.2. Gradient Judgments and Non-Categorical Grammars

This study exclusively deals with grammars which are categorical. BUFIA, as described and implemented here, is categorical by default, and W&G require MaxEnt-Ftr to induce only surface-true constraints. This is to provide faithful comparison across model types and because the Bolivian Quechua phonotactics under consideration are argued by W&G to be categorical in nature. However, this raises questions about whether an investigation of the role of statistics in phonotactic learning is complete without considering phonotactic phenomena which exhibit gradient judgment effects.

Hayes and Wilson (2008) (a.o.) argue that probabilistic grammars are necessary to account for gradience in acceptability judgments. Some recent studies dispute this, however, showing that categorical grammars match or outperform probabilistic grammars in terms of correlation with native speaker acceptability data (Gorman, 2013; Kostyszyn and Heinz, 2021). Durvasula (2020, 2025) argued that the presence of type frequencies does not actually improve the performance of the Hayes and Wilson (2008) MaxEnt learner on English onset cluster data. In light of these findings, it seems appropriate to leave gradient judgments out of scope for this study.

That said, it is possible to include constraint weights within BUFIA's structural learning approach. One way to modify BUFIA to return weighted constraints would be to make it sensitive not just to the presence or absence of a certain structure in the input data, but also to its frequency. In fact, as explained next, BUFIA offers a framework in which questions about which properties of phonotactic learning algorithms are necessary and/or sufficient can be explored clearly and directly.

4.3. BUFIA as a General-Form Learning Framework

Although we have so far described a single, monolithic version of BUFIA, in reality BUFIA is not a single algorithm but a class of algorithms, all with the same basic procedure described in section 2.2, one instantiation of which is used in this paper. Each BUFIA variant makes different choices regarding the traversal of the constraint space. These choices are not directly data-dependent, and constitute explicit and independently adjustable hyperparameters of the model.[4] These hyperparameters include the stopping conditions, constraint selection criteria, constraint ordering (between factors on the same "layer"), and other representational choices. The stopping condition (as discussed in Section 3.2) was explicitly tuned in our experiments and corresponds to the maximum number and/or maximum size of constraint that the model is willing to consider. Constraint selection criteria refers to how the model decides whether or not to add an identified constraint to the grammar. In the version of BUFIA used in our experiments, we add new constraints to the grammar if and only if they rule out some form(s) not already banned by the constraints compiled thus far. A (hypothetical) gradient version of BUFIA, however, could employ an alternate strategy for constraint selection that was sensitive to the actual and expected frequencies of structures in the input sample. In our model, features which pick out larger natural classes are preferred, which impacts the order in which constraints on the same "layer" of the constraint hierarchy will be considered.

Rawski (2021) explores the consequences of some different strategies for constraint ordering and constraint selection, and the implications for phonotactic learning as abductive–as opposed to inductive–inference.

Finally, one of BUFIA's biggest strengths is that it is completely agnostic to the choice of representation. In this paper we use sequences of feature bundles, ordered either segmentally or over a tier projection. However, versions of BUFIA have been developed for learning tonotactic constraints where the factors correspond to autosegmental representations (Li, 2025), as well as for comparing the phonotactic learning of grammars containing "banned" subfactors vs "allowed" subfactors (Payne, 2024).

Because these hyperparameters are modular–i.e., each can be modified independently of the others–and entirely explicit, BUFIA as a class of algorithms offers an ideal setup to conduct ablation studies in which to carefully probe the impact of the aforementioned different choices on learning behavior. Specific research questions, like the one discussed at the end of Section 4.2, can be addressed by comparing versions of BUFIA which differ only by the principle in question: a sandbox of sorts for running experiments on learning.

5. Conclusion

W&G use the strong performance of MaxEnt-Ftr in their initial experiment to argue that both *features* and *statistics* are necessary ingredients for successful phonotactic learning. We demonstrate here that the deterministic, structure-based learning algorithm BUFIA performs competitively with MaxEnt-Ftr, both under the same basic learning setup used by W&G, and an alternate setup designed to eliminate potential confounds. In light of these results, we argue that while features do seem to be important for phonotactic learning, statistics need not be. Rather, the internal structure of the hypothesis space plays a vital role, which can be exploited by the learner to search large spaces in a principled way. Learning without statistics, far from being doomed, is a promising approach which

lends deep insight into the nature of the learning problem and the factors which influence learning behavior.

References

Ajacopa, Teofilo Laime, Efraín Cazazola, Félix Layme Pairumani, and Pedro Plaza Martínez. 2007. Diccionario bilingüe iskay simipi yuyayk'ancha. *La Paz, 2007 (Quechua)* .

Chandlee, Jane, Remi Eyraud, Jeffrey Heinz, Adam Jardine, and Jonathan Rawski. 2019. Learning with partially ordered representations. In *Proceedings of the 16th Meeting on the Mathematics of Language*, edited by Philippe de Groote, Frank Drewes, and Gerald Penn, 91–101. Toronto, Canada.
URL https://aclanthology.org/W19-5708

Dolatian, Hossep. 2020. Computational locality of cyclic phonology in Armenian. Doctoral dissertation, Stony Brook University.

Durvasula, Karthik. 2020. O gradience, whence do you come? Keynote presentation at the annual meeting of phonology.

Durvasula, Karthik. 2025. A closer look at what/how we can learn from computational modelling of phonotactics. *Lingbuzz Preprint* .

Gallagher, Gillian. 2011. Acoustic and articulatory features in phonology–the case for [long VOT]. *The Linguistic Review* 28:281–313.

Gallagher, Gillian. 2016. Vowel height allophony and dorsal place contrasts in cochabamba quechua. *Phonetica* 73:101–119.

Gorman, Kyle. 2013. Generative phonotactics. Doctoral dissertation, University of Pennsylvania.

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

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. 2023. Statistical learning. In *An introduction to statistical learning: With applications in Python*, 15–67. Springer.

Jardine, Adam. 2017. The local nature of tone-association patterns. *Phonology* 34:385–405.

Kostyszyn, Kalina, and Jeffrey Heinz. 2021. Categorical account of gradient acceptability of word-initial polish onsets. In *Proceedings of the Annual Meetings on Phonology*.

Lambert, Dakotah. to appear. Multitier phonotactics. *Phonology* .

Li, Han. 2025. Learning tonotactic patterns over autosegmental representations. In *Proceedings of the Annual Meetings on Phonology*, vol. 1. University of Massachusetts Amherst Libraries.

Payne, Sarah. 2024. A generalized algorithm for learning positive and negative grammars with unconventional string models. In *Proceedings of the Society for Computation in Linguistics 2024*, edited by Richard Futrell, Connor Mayer, and Noga Zaslavsky, 75–85. Irvine, CA: Association for Computational Linguistics.
URL `https://aclanthology.org/2024.scil-1.8/`

Rawski, Jonathan. 2021. Structure and learning in natural language. Doctoral dissertation, Stony Brook University.

Rogers, James, and Dakotah Lambert. 2019. Some classes of sets of structures definable without quantifiers. In *Proceedings of the 16th Meeting on the Mathematics of*

Language, 63–77. Toronto, Canada: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-5706

Strother-Garcia, Kristina. 2018. Imdlawn Tashlhiyt Berber syllabification is quantifier-free. In *Proceedings of the Society for Computation in Linguistics*, vol. 1. Article 16.

Torero, Alfredo. 1964. *Los dialectos quechuas*. Univ. Agraria.

Wilson, Colin, and Gillian Gallagher. 2018. Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic Inquiry* 49:610–623.

Logan Swanson (Stony Brook University): logan.swanson@stonybrook.edu
Jeffrey Heinz (Stony Brook University): jeffrey.heinz@stonybrook.edu
Jonathan Rawski (San Jose State University): jon.rawski@sjsu.edu

Notes

We gratefully acknowledge support from the Ookami community under US NSF grant #1927880. This research was also supported by US NSF grant #2416183 to JH. We also thank Colin Wilson and Gillian Gallagher for help with their materials, and Jordan Kodner, Ellen Broselow, and audiences at AMP, MIT, and Rutgers for helpful feedback.

[1]We avoid the terms ‘substructure’ and ‘superstructures’ because they have related, but different, meanings in model theory.

[2]We have described the constraint space as growing upwards. However, this is purely by convention–it could just as easily be described as rooted at the “top” and growing downwards, or growing left to right.

[3]Our implementation of BUFIA can be found here:

https://github.com/pterodactylogan/bufia/tree/ling_inquiry.

[4]Additional variants and their description are available in the Github codebase.

	Statistics	**No Statistics**
Segments	MaxEnt-Seg	(T)SL
Features	MaxEnt-Ftr	?

Table 1: Paradigm of model types laid out by W&G, with each model varying on representational choice and use of statistics. (T)SL abbreviates (Tier) Strictly Local.

	i	u	e	o
high	+	+	−	−
back	−	+	−	+

Table 2: A simple toy vowel system

ii	ie	ei	ee
uu	uo	ou	oo

Table 3: Strings of length 2 exhibiting backness harmony

Tier	Segments	Phonotactic generalizations
dorsal	dorsal consonants, vowels	high-mid vowel allophony
c-dorsal	dorsal consonants, morpheme boundary	no co-occurrence of velars and uvulars within morphemes
laryngeal	h, ʔ, stops, affricates	restrictions on aspirates and ejectives
segmental	all	syllable structure, *[wu], *[wo]

Table 4: Summary of phonotactic generalizations in Quechua, as described and grouped into tiers by Wilson and Gallagher (2018)

	held-out forms (W&G)	legal nonce roots	illegal nonce roots
Features, Stats (MaxEnt-Ftr)	100%	96.4%	6.5%
Features, No-stats (BUFIA)	99.6%	94.1%	1.8%
Segments, Stats (MaxEnt-Seg)	100%	98.9%	58.8%
Segments, No-stats ((T)SL)	95.7%	18.1%	0.004%

Table 5: Experiment 1 Results: Percentage of forms accepted by evaluation category aggregated over the five folds. All results correspond to performance on the evaluation set.

	Precision	Recall	F1	Held-Out	Illicit
Features, Stats (MaxEnt-Ftr)	0.769	0.991	0.865	99.1%	30.1%
Features, No-stats (BUFIA)	0.946	0.943	0.945	94.3%	5.05%
Segments, Stats (MaxEnt-Seg)	0.554	1.0	0.713	100%	79.8%
Segments, No-stats ((T)SL)	1.0	0.429	0.600	42.9%	0.0%

Table 6: Experiment 2 results, with scores aggregated over the five folds. Percentages represent the number of forms accepted by category.

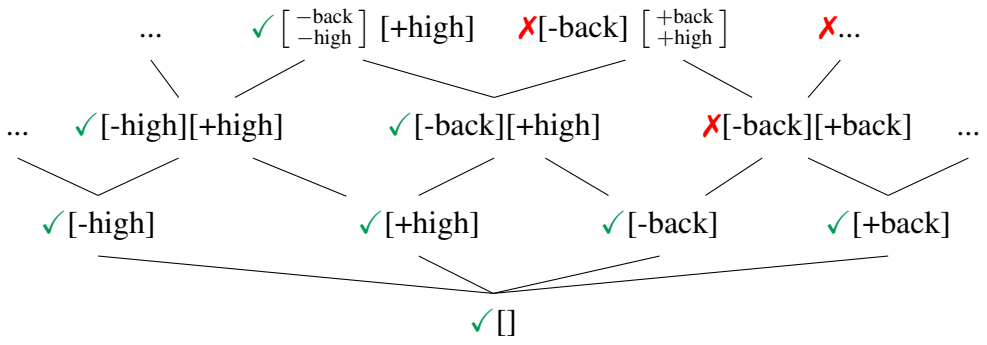

Figure 1: Fragment of the hierarchy of the space of possible constraints, with superfactors dominating their subfactors. If a structure is illicit, it is a valid constraint and everything above it must be illicit. If a structure is licit, it is not a constraint and everything below it must be licit.

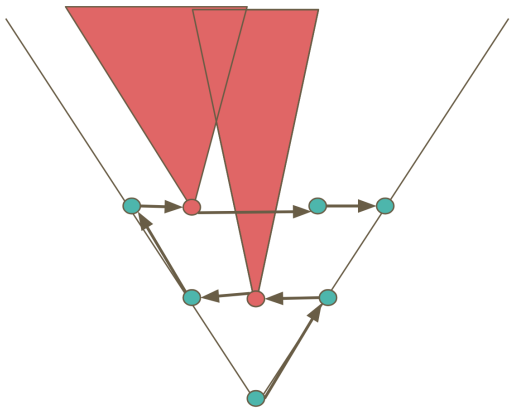

Figure 2: The BUFIA algorithm traversing and pruning the search space. Arrows indicate direction of search, while shaded areas indicate sections of the partial order which have been pruned.