Phonotactic Learning and Constraint Selection Without Statistics

Logan Swanson[1], Jeffrey Heinz[1], and Jonathan Rawski[2]

[1]Stony Brook University, [2]San Jose State University

Ingredients for Learning

A core problem in phonotactic learning is deciding ***which* constraints to add to the grammar**. The **MaxEnt learner**[1] decides in part by statistics and in part by other heuristics and parameters of the model.

Wilson and Gallagher[2] explored the role of statistics and representational choice (feature-based or segment-based) on learning by evaluating different phonotactic learning models using phonotactic constraints in Bolivian Quechua.

They compare three models:

1. a feature-based MaxEnt model
2. a segment-based MaxEnt model
3. a segmental-tier-based non-statistical model

W&G conclude that both statistics and featural representations are necessary for learning.

What about a non-statistical feature-based model?

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

*"What about a nonstatistical model that learns by memorizing feature sequences?... Lacking a method for deciding **which** representations are relevant for assessing well-formedness – precisely the role played by statistics in Maxent-Ftr – **learning... is doomed**."*

– Wilson and Gallagher 2018

Completing the Paradigm

BUFIA[3] (Bottom-Up Factor Inference Algorithm), a deterministic abductive algorithm which leverages the structure of the constraint space to find general constraints, is exactly this fourth type of non-statistical feature-based learner.

We demonstrate that BUFIA can perform as well or better than the feature-based MaxEnt model on the same data, showing that **statistics are not necessary for phonotactic learning.**

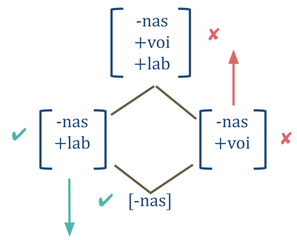
A Non-Statistical Learner

Structure of the Constraint Space

If constraints are represented by sequences of feature bundles, the exponential combination of features yields a massive search space.

However, this space is not a random grab-bag, but rather has a rich internal structure, forming a dense partial order.

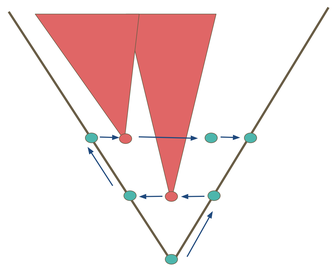
- **Licitness proceeds downwards:** If non-nasal labial segments are permitted, then non-nasal segments must be permitted
- **Illicitness proceeds upwards:** If non-nasal voiced segments are illicit, then non-nasal voiced labial segments must be illicit

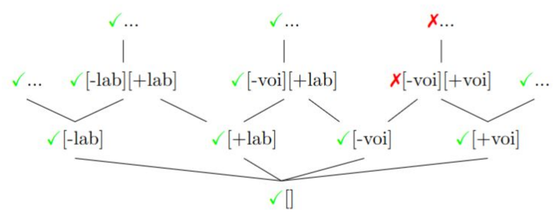


The BUFIA Algorithm

The Bottom-Up Factor Inference Algorithm (BUFIA)[3] is able to learn feature-based grammars using a breadth-first search of the partial order formed by the possible constraints and banning any structures not present in the input.

- Start from the bottom of the partial order, and proceed upwards breadth-first
- For each factor:
  - **If it is present in the data, continue**
  - **If it is not present in the data, add it as a constraint and prune the search space**
- Stop when a cutoff condition is reached:
  - Typically this will be an upper bound on the size of the factors
- The output will be a list of constraints, ordered by generality





BUFIA traversing an abridged phonotactic search space. The most general constraints are checked first.

Try the code!

Experimental Results

Experiment 1

We duplicated the training setup used by Wilson & Gallagher[2].

Training data: 1,000 dictionary forms appended with one of three possible suffixes (or no suffix). These were divided into five folds, with 20% held out in each fold.

Testing data: In addition to the held-out forms, testing included all possible CV(C)CV(C) sequences, classified as "legal" (150k) and "illegal" (400k), according to known phonotactic generalizations.

	held-out forms (W&G)	legal nonce roots	illegal nonce roots
Features, Stats (MaxEnt-Ftr)	99.8%	82.2%	1.9%
Segments, Stats (MaxEnt-Seg)	99.7%	71.5%	45.4%
Segments, No-stats ((T)SL)	96.7%	18.8%	0.1%
Features, No-stats (BUFIA)	99.6%	94.1%	1.8%

Table 1: Experiment 1 Test Results: Percentage of forms accepted by evaluation category aggregated over the five folds. Results reported in rows 1-3 are from W&G.

Experiment 2

- In Experiment 1, roots are duplicated 4x in the training set, and not controlled for in fold construction
  → many roots are present in both train and test sets.
- The distribution of illicit forms which violate each known constraint is highly skewed
  → constraints on one tier (out of 4) able to rule out 89% of illicit forms.
- Synthetic "licit" data is unverified by native speakers, baking in an assumption that the constraints identified by W&G are the only ones active in the grammar.

To ensure none of these factors artificially boosted model performance, we conducted a second experiment:

Training Data:

- Alternate 5-fold split of dictionary forms
- no roots duplicated across train and test sets

Testing Data:

- Held-out dictionary forms
- 40 forms which uniquely violate each known constraint
- Balanced quantity of held-out and illicit forms (allows for reporting f1)

	precision	recall	f1
MaxEnt-Ftr	0.786	0.951	0.861
BUFIA	0.946	0.943	0.945

Table 2: Experiment 2 Test Results: Scores are aggregated over the five folds.

References: [1] Hayes, B. and C. Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39. [2] Wilson, C. and G. Gallagher (2018). Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic Inquiry 49*. [3] Chandlee, J., Eyraud, R., Heinz, J., Jardine, A., & Rawski, J. (2019). Learning with partially ordered representations.