

## Phonology vs. phonotactics

- The English prefix *in-*, meaning “not”, takes a different form depending on the place and manner of the following consonant: [ɪ] (e.g., *i[m]ature*), [ɪm] (e.g. *i[m.p]ossible*), [ɪn] (e.g. *i[n.t]angible*), and [ɪŋ] (e.g. *i[ŋ.g]ratitude*)
- Recent trends in phonology (associated with Optimality Theory) emphasize formal constraints on sound sequences over rules (i.e., functions) as the triggers of such phonological *alternations*

• *Constraint:* Disprefer [NASAL, α PLACE] [β PLACE] iff α ≠ β

• *Rule:* [NASAL] → [α PLACE] / \_\_\_ [α PLACE]

## The null hypothesis

- Halle [1] argues that there is no need for *phonotactic constraints* and that any unattested sequence is either ruled out by a rule (as above), or an *accidental gap* in the language’s lexicon
- Others argue that constraints act in the lexicon, disfavoring certain sound sequences in words, regardless of whether these constraints correspond to that language’s alternations [4]
- Lexical gaps or dispreferences are insufficient to disprove Halle’s null hypothesis; rather, it is necessary to further show that these dispreferences do not correspond to alternations in the language

## A syllable contact corpus study

- To investigate lexical dispreferences in English, we collected all 5,810 monomorphemic *syllable contact clusters* (sequences of one or more word-internal onset consonants immediately preceded by one or more coda consonants) in the CELEX database (data available upon request)
- There are 48 onset types and 21 coda tokens, but only 131 clusters out of 1,008 (= 48 × 21) possible clusters (*saturation rate* 13%)

## The role of phonology

- English phonological alternations [2] account for much of the missing data:
  - *Nasal assimilation* (see left panel) rules out [m.t], [n.p]... (but *da[m.z]el*)
  - *Voice assimilation* (*wal[kt]* vs. *dra[gd]*) rules out [gt], [kd]... (but *vo[d.k]a*)
  - *Degemination* (*i[m]ature* vs. *i[m.b]alance*) rules out [m.m], [t.t]...
- All these alternations are reliable predictors of which clusters are and are not attested in the database (Fisher exact test  $p < 0.05$ )
- These alternation-based generalizations raise the saturation rate to 21%

## The role of phonotactics

- Pierrehumbert [4] proposes three constraints on syllable contact clusters which lack any corresponding alternation:
  - \**Dorsal-labial* (but *do[gm]a*, *eni[gm]a*, etc.)
  - \**Coda coronal obstruent* (but *a[t.l]as*, *no[s.tr]il*, etc.)
  - \**A.BA* (“geminate with intervening consonant”)
- Due to the many exceptions to these generalizations, we fail to reject the null hypothesis that any under-attestation conforming to these constraints is due to chance (Fisher exact test  $p > 0.05$ )

## The role of data sparsity

- By the *Turing estimate* ( $p_0 = n_1 / N$ ) [3], 3% (= 32 / 1174) of the probability should be reserved for unseen clusters, and thus many unattested clusters might be missing due to sampling
- The data is an excellent fit ( $R^2 = 0.91$ ) to the *Generalized Zipf’s Law*

$$f(r; \alpha) = C / r^\alpha$$
 where  $f(r)$  is the frequency of the  $r$ th outcome (see figure to right)

## A phonotactic modeling bake-off

- Each computational model is trained on attested clusters, and assigns a wellformedness score to each of the 1,008 possible clusters
- A “stump classifier” (a soft linear-kernel SVM) converts this continuous signal to a binary classification (*attested* vs. *unattested*)

Model	Accuracy	Precision	Recall	F-score
Baseline (“all unattested”)	0.87	(n.a.)	(n.a.)	(n.a.)
Alternations (“neutralizations unattested”)	0.90	0.27	0.75	0.40
Segment-based joint MLE probability [5]	0.87	0.08	0.59	0.14
Feature-based MaxEnt probability [6]	0.93	0.69	0.80	0.74

## Conclusions

Phonotactic learning (whether by infant or computer) suffers from an acute *data sparsity problem*. By ignoring this, the phonotactic modeling literature conflates *structural* and *accidental gaps* in the lexicon, and fails to credit sparsity and alternation as possible causes of gaps

