Grammars Leak:

How categorical phonotactics can cause gradient phonotactics

Andy Martin amartin@humnet.ucla.edu **UCLA**

Questions

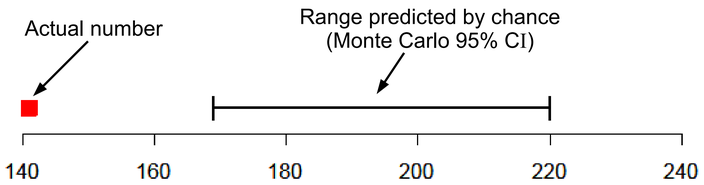
- What are the effects of phonotactics on morphological operations?
- How are tautomorphemic and heteromorphemic phonotactics related?
- Under what conditions do learners sacrifice accuracy for simplicity?

Data

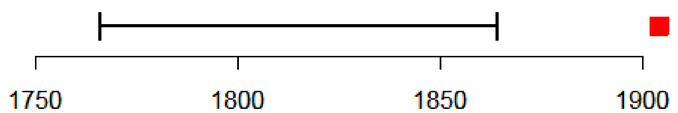
- Geminates are only allowed in English across morpheme boundaries:

  boo[kk]*ase, sou*[ll]*ess*
  carpool versus *carp pool*

- But in compounds, fewer geminates occur than are predicted by chance
- Number of CELEX noun-noun compounds with geminates (out of 4,578):



Compare to legal CC clusters across compound boundary:



- Geminates are legal in compounds, but underrepresented

Other cases:

- Navajo compounds tend to obey sibilant harmony
- Turkish compounds tend to obey vowel harmony

Categorical phonotactics within morphemes are gradiently obeyed across morpheme boundaries

Hypothesis

Phonotactics "leak" from tautomorphemic into heteromorphemic domain for two reasons:

- **The presence in the grammar of constraints that are blind to morphological structure**
- **A learning bias in favor of simpler grammars**

- The phonotactic grammar is modeled using weighted markedness constraints and a Maximum Entropy learning algorithm (see box below)
- Strategy: train learner on tauto- and heteromorphemic consonant clusters and show that it learns a gradient phonotactic even when the data is not biased
- The training data consists of biconsonantal clusters of [p] and [t], with an optional morpheme boundary:

Cluster	Structure	Number of examples
pt	monomorpheme	2000
tp	monomorpheme	2000
p+t	compound	1000
t+p	compound	1000
p+p	compound	1000
t+t	compound	1000

No bias in training data

- Tautomorphemic geminates [pp], [tt] do not occur in training data, but heteromorphemic geminates occur freely

Maximum Entropy Grammars

- Grammar consists of a set of OT-like constraints
- Each constraint has non-negative real number weight
- Candidates are assigned a *score*: the sum of (weight * violations) for every constraint:

	C_1 (w 0.5)	C_2 (w 1.0)	C_3 (w 2.0)	$\Phi(x)$ (*score*)
x	** $2 \times 0.5 = \mathbf{1}$	$0 \times 1.0 = \mathbf{0}$	* $1 \times 2.0 = \mathbf{2}$	$1 + 0 + 2 = \mathbf{3}$

- The score can be used to compute the *probability* of the candidate (higher score = lower probability)
- Learning algorithm finds the grammar that maximizes the probability of the data
- Algorithm also includes smoothing term:

$$\sum_{i=1}^{N} \log P(x_i) - \sum_{j=1}^{M} \frac{w_j^2}{2\sigma_j^2}$$

Probability of data — Smoothing term

The smoothing term penalizes high constraint weights. This is necessary to avoid overfitting the training data.

- The learning algorithm was run twice: first, using only constraints that are sensitive to morphological structure:

Structure-sensitive constraints:

***pp**	no geminates within morpheme
***tp**	no non-geminate clusters within morpheme
***p+p**	no geminates across morpheme boundary
***t+p**	no non-geminate clusters across morpheme boundary

Grammar learned with structure-sensitive constraints

Constraint weights	
*pp: 4.02	*p+p: 0
*tp: 0.12	*t+p: 0

No bias against heteromorphemic geminates

- Next, the learner was run again on the same data—this time, constraints that ignore morphological structure were added to the structure-sensitive constraints:

Structure-blind constraints:

***p(+)p**	no geminates
***t(+)p**	no non-geminate clusters

- Note that *p(+)p is violated less often in the training data, simply because pp does not occur

Grammar learned with both constraint types

Structure-sensitive		Structure-blind	
*pp: 4.01	*p+p: 0	*p(+)p: 0.03	
*tp: 0.13	*t+p: 0	*t(+)p: 0	

Slight bias against heteromorphemic geminates

Why does this happen?

- The smoothing term in the learning algorithm introduces a tradeoff between maximizing the probability of the data (accuracy) and giving constraints low weights (complexity)
- Giving *p(+)p a nonzero weight reduces the accuracy of the grammar, since it predicts fewer p+p than t+p
- This reduces the penalty incurred for high weights, since it allows the weight of *pp to be decreased—the work of explaining why [pp] is unattested is shared between *pp and *p(+)p

Conclusions

- **Root-internal phonotactics can have gradient effects on morphological processes**
- **This process can be modeled as a side-effect of the learner's bias against complex grammars**