Grammars leak:
How categorical phonotactics can cause gradient phonotactics

In this poster I present evidence from Turkish, Navajo, and English that categorical phonotactics that hold within morphemes can be accompanied by gradient versions of the same phonotactics in heteromorphemic contexts. In English, for example, geminate consonants are forbidden within morphemes, but permitted across morpheme boundaries, e.g., *taillight*, *bookcase*. English words with geminates, however, although legal, are statistically underrepresented. Of the 5,070 noun-noun compounds in the CELEX database (Baayen et al. 1993), 3.0% contain geminates, significantly lower than the rate of 4.1% that would be expected by chance ($p<.0001$ by Monte Carlo test). In other words, a sound sequence that is banned within roots is legal, but dispreferred, across morpheme boundaries. I show that the same is true for vowel harmony in Turkish (compounds with disagreeing vowels are underrepresented) and sibilant harmony in Navajo (compounds with disagreeing sibilants are underrepresented).

I argue that these gradient phonotactics are the result of the presence of two constraint types in the phonotactic grammar: those that are sensitive to morphological structure (e.g., a constraint banning geminates within morphemes), and those that are blind to structure (e.g., a constraint that simply bans all geminates). I model the English data with a Maximum Entropy ("maxent"; Della Pietra et al. 1997, Goldwater and Johnson 2003, Hayes and Wilson to appear) learning algorithm which uses both structure-sensitive and structure-blind constraints.

A maxent phonotactic grammar consists of a set of markedness constraints, each of which has a weight, represented by a nonnegative real number. The grammar maps candidate outputs to probabilities; candidates that violate constraints with high weights have correspondingly lower probabilities. The learning algorithm, when given a set of constraints and a set of training data, adjusts the weights of the constraints until the expected number of constraint violations equals the actual number of violations, which is equivalent to maximizing the probability of the data (the learner finds the maximum of the function in Figure 1).

Because simply maximizing the probability of the data can result in overfitting the data, however, maxent learning is typically performed in conjunction with a smoothing method, which in my model consists of a Gaussian prior on the constraint weights (Chen and Rosenfeld 2000). The prior penalizes high weights (I set the ideal weight μ for all constraints to zero) and thus exerts a pressure to generalize, even at the cost of accuracy in modeling the training data. I gave the maxent learner both structure-sensitive and structure-blind constraints and exposed it to data in which geminates do not occur within morphemes, but occur freely across morpheme boundaries. Predictably, it assigns a high weight (4.01) to the structure-sensitive constraint against tautomorphemic geminates, correctly ruling such geminates out as illegal. But the prior also forces the learner to assign a small, nonzero weight (0.03) to the structure-blind constraint against all geminates, despite the fact that there is no bias against heteromorphemic geminates in the training data. The algorithm is in effect predicting that heteromorphemic geminates should be slightly ill-formed based on the absence of tautomorphemic geminates. The pressure to generalize enforced by the Gaussian prior can thus result in overgeneralization, causing a phonotactic in a restricted domain to "leak" into other domains. These results have consequences for theories of phonotactic learning, phonotactic restrictions on morphological operations, and the interaction between categorical and gradient generalizations.

Figure 1. Function maximized by the learning algorithm

$$\underbrace{\boxed{\sum_{i=1}^{N} \log P(x_i)}}_{\text{probability of the data}} - \underbrace{\boxed{\sum_{j=1}^{M} \frac{(w_j - \mu_j)^2}{2\sigma_j^2}}}_{\text{Gaussian prior}}$$

$\langle x_1, x_2 \ldots x_N \rangle$:	outputs in training data
$P(x_i)$:	probability of output x_i
$\langle w_1, w_2 \ldots w_M \rangle$:	constraint weights
$\langle \mu_1, \mu_2 \ldots \mu_M \rangle$:	priors on constraints (set to zero for all constraints)
σ^2:	free parameter

References

Baayen, R.H., R. Piepenbrock, and van H. Rijn. 1993. *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium. University of Pennsylvania, Philadelphia, PA.

Chen, Stanley and Ronald Rosenfeld. 2000. A Survey of Smoothing Techniques for ME models. *IIEE Transactions on Speech and Audio Processing* 8: 37-50.

Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19: 380–393.

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Osten Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theor*y. 111–120.

Hayes, Bruce and Colin Wilson. To appear. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*.