Evaluating distance-based alternatives to natural class-based constraints

Blake H. Allen
Department of Linguistics, University of British Columbia

Dilemma for Phonology

Natural classes are insufficiently expressive. What alternatives do we have?

- Without segment classes, generalization to novel word-shapes is impossible (e.g. failing the "Bach test").
- But allowing a class for every possible segment combination creates a massive burden for learnability (2^n classes for an inventory of n segments).
- Natural classes have been the standard solution to this problem: generalization is possible, and the search space is small(er), plus empirical coverage is good.
- However: humans make phonological generalizations unbounded by natural classes (Cristia et al. 2013)!

⇒ **What segment classes are consistent with Cristia et al.'s (2013) findings, are reasonably learnable, and can also achieve natural class-like generalizations?**

Motivation: Cristia et al. (2013)

Their research question: when humans get evidence that a particular set of segments is licit in a particular position, which other segments will they deem also licit there?

Their two relevant hypotheses:

1. Generalize to missing members of the smallest natural class containing observed segments (e.g., below, to only/mostly [b]).
2. Generalize to segments featurally similar to observed ones (e.g., below, to [b] and [k] equally).

k
↑
b ← d g v z ʒ

N.B.: This poster uses Cristia et al.'s (2013) findings as a conceptual jumping-off point; it is not an attempt to directly implement their featural distance hypothesis.

Quasi-clique constraints

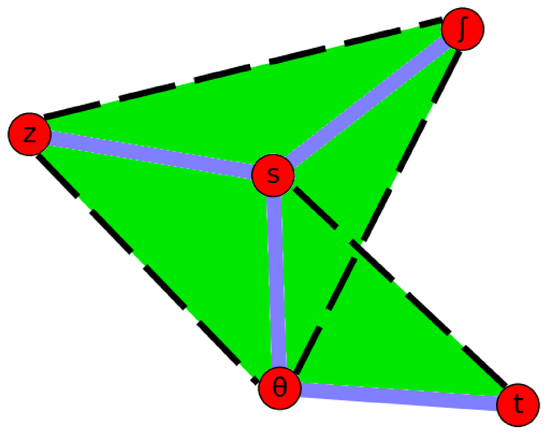
Quasi-clique: a set of featurally similar segments which can behave together a class.

Definition:
A set S of segments is a quasi-clique if and only if, for every segment s in the set, every other segment in the set differs from s in fewer than $|S|$ feature values.
• Based on, but not identical to: Matsuda et al. (1999), Brunato et al. (2008).

Not all quasi-cliques are natural classes, e.g. [ð|ʒ|s].

This constraint family successfully predicts the Cristia et al. (2013) results when positive constraint weights are allowed (more on methodology in rightmost column).

Examples of quasi-cliques



Red nodes: 1-segment quasi-cliques
Blue edges: 2-segment quasi-cliques
Green triangles surrounded by two blue edges and one dotted line: 3-segment quasi-cliques
Not indicated: various 4-segment quasi-cliques and one 5-segment clique (the whole graph)

Resulting unigrams: [z], [s], [ʃ], [θ], [t],
[z|s], [s|ʃ], [s|θ], [θ|t]
[z|s|ʃ], [z|s|θ], [s|ʃ|θ], [s|θ|t]...

Test case: English onsets

Quasi-clique constraints allow generalization—including that seen in Cristia et al.'s (2013) results—and limit the search space enough that learning should be possible.

But how do they perform on real language data?

Methods

Dataset: the English onsets data from Hayes & Wilson (2008), i.e. from the CMU Pronouncing Dictionary
Features: same as in Hayes & Wilson (2008)
Quasi-clique parameters:
- maximum quasi-clique size: 3
- augmented with wildcard unigram (any segment)
- maximum constraint size: bigram

Weight optimization: phonotactic learning module in PhoMEnt (Daland et al. 2014); all weights ≤ 0
Testing data: same as in Hayes & Wilson (2008)

Results

Spearman's correlations between model predictions and experimental (wug) data were found using R's *cor* function.

- Hayes & Wilson (2008) constraints: $\rho = 0.887$ (comparable to the 0.889 originally reported)
- Quasi-cliques: $\rho = 0.847$
- An ANOVA (R: anova) shows that the superset model (both grammars together) is more predictive than just the natural class-based model ($p < 0.05$).

⇒ The quasi-clique constraints capture learned generalizations not expressed by the natural class constraints.

Acknowledgments

Special thanks to Robert Daland and Kevin McMullin.

(References are available here at this station or digitally by request.)