Generalization despite variation: French schwa with lexically indexed constraints

Aleksei Nazarov, Utrecht University, a.i.nazarov@uu.nl
Brian Smith, Independent Researcher, bwsmith.linguist@gmail.com

In modeling lexically specific phonological phenomena, there is a tradeoff between underfitting and overfitting; if a model better accounts for the differences between individual words, it will generalize less well to novel forms, and *vice versa*. Balancing this tradeoff is especially challenging for Optimality Theoretic models that use lexically indexed constraints (constraints that apply only to certain words/morphemes; e.g., Kraska-Szlenk 1995, Pater 2000) when the learning algorithm can create new indexed constraints (e.g., Becker 2009, Round 2017), or cophonologies (constraint rankings that apply only to certain words/morphemes; Inkelas & Zoll 2007) when the learning algorithm can create new grammars. In this paper, we examine this tradeoff for French schwa deletion by evaluating three models of constraint indexation that differ in how many rounds of grammar optimization are needed. We show that increasingly more complex models yield an increasingly better fit to data, while still maintaining generalization to new words.

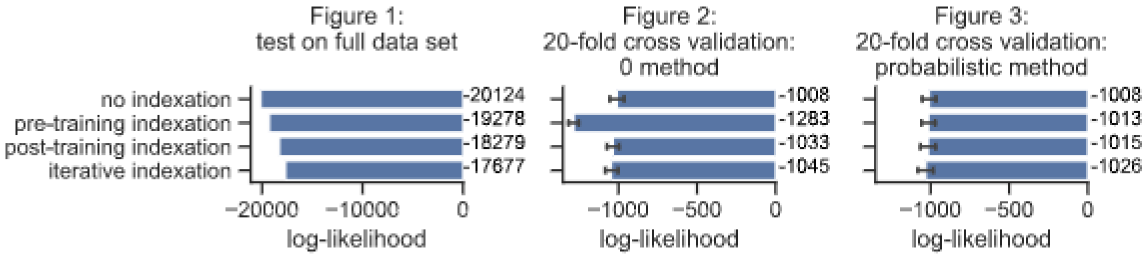
We consider four MaxEnt-based (Goldwater & Johnson 2003) models, which differ in how they determine lexically indexed constraints. The process of creating indexed constraints and connecting them to lexical items has not been implemented in MaxEnt grammars before, so a variety of models should be tested. The **no indexation** model only has non-indexed constraints whose weights are learned in a single round of optimization (using the hgR MaxEnt learning software, Staubs 2011). The **pre-training** model uses constraint violations and candidate frequencies to estimate whether a word should be indexed to a constraint. This occurs before optimization, meaning that constraint weights and interactions between constraints don't play a role in indexation. The **post-training** model first trains a model without indexed constraints, and then uses information from that model to determine which words to index to a set of indexed constraints, after which the full model is trained again. Finally, the **iterative** indexation model is trained multiple times, in turn adding one best-fitting indexed constraint, indexing a corresponding set of best-fitting words to it, and optimizing the weights. This is repeated until the learner stops improving. While the pre- and post-training model have a fixed set of indexed constraints (each non-indexed constraint has an indexed version), the iterative model generates a set of indexed constraints that varies depending on the patterns in the dataset. We also consider two ways to generalize the grammar to novel forms (i.e., forms outside the training data): the **0 method** (no indexed constraint violations for novel inputs), and a **probabilistic method** (violations for new inputs are proportional to the number of training inputs indexed to that constraint; cf. Becker 2009).

To compare these models, we use French 'schwa' deletion (e.g., Dell 1985), a probabilistic phonological process which is mediated by both lexical and phonological factors. We used deletion frequencies for 456 words based on judgment experiments from Racine (2008). The examples in (1) show how deletion rates can differ between words with schwa in very similar phonological environments.

(1) /sə̠mɛn/ 'week' (50% deletion)   /sə̠mɛstʁ/ 'semester' (14% deletion)

For these words, tableaux were constructed with a simple constraint set based on Kaplan's (2011) analysis of French schwa: *ə, *ə/¬[_.σ#] (pro-deletion constraints), Max, *#CC, *CCC, *CC$_{[+nas]}$C, *CC$_{[-son,-cont]}$C$_{[+nas]}$ (anti-deletion constraints).

To evaluate the degree of underfitting, we trained and tested each of the four models on the full dataset. For overfitting, we used 20-fold cross-validation, dividing the data set into 20 chunks, with 19 chunks to train the model and the remaining chunk to evaluate it (repeating 20 times in total), and then averaging performance across the 20 runs. Model performance is shown below in terms of data log-likelihood, which indicates the probability of the data given the model in log-space (closer to zero is better), with 95% CIs. The scale is different between Figures 1 and the other two because log-likelihood depends on the size of the data set.



As shown in Figure 1, the likelihood of the full dataset goes up as model complexity increases (moving down the graph). For the cross-validation results in Figures 2 and 3, all of the models seem to generalize equally well (except for the pre-training model in Figure 2). Thus, while the models (save one) cannot be distinguished by generalization, the more sophisticated the indexation procedure, the better lexically specific patterns are captured. We conclude that (machine-determined) constraint indexation can indeed provide a better fit to phonological data while still retaining the ability to generalize to new forms. Even adding the simplest type of indexation helps account for variation and does not hurt generalization.

References

Becker, M. 2009. *Phonological Trends in the Lexicon: The Role of Constraints.* University of Massachusetts Amherst dissertation. ◆ Dell, F. 1985. *Les règles et les sons: Introduction à la phonologie générative*. Paris: Hermann. ◆ Goldwater, S. & M. Johnson. 2003. Learning OT Constraint Rankings Using a Maximum Entropy Model. In J. Spenader, A. Eriksson, Ö. Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–20. ◆ Inkelas, S. & C. Zoll. 2007. Is Grammar Dependence Real? A Comparison Between Cophonological and Indexed Constraint Approaches to Morphologically Conditioned Phonology. *Linguistics* 45.1:133–71. ◆ Kaplan, A. 2011. Variation through Markedness Suppression. *Phonology* 28.3:331–370. ◆ Kraska-Szlenk, I. 1995. *The Phonology of Stress in Polish*. University of Illinois at Urbana-Champaign dissertation. ◆ Pater, J. 2000. Non-uniformity in English secondary stress: The role of ranked and lexically specific constraints. *Phonology* 17.2:237–74. ◆ Racine, I. 2008. *Les effets de l'effacement du Schwa sur la production et la perception de la parole en français*. Université de Genève dissertation. ◆ Round, E. 2017. Phonological exceptionality is localized to phonological elements: The argument from learnability and Yidiny word-final deletion. In C. Bowern, L. Horn & R. Zanuttini (eds.), *On looking into words (and beyond): Structures, relations, analyses*, 59–97. Berlin: Language Science Press. ◆ Staubs, R. 2011. *Harmonic Grammar in R (hgR).* Software package. http://blogs.umass.edu/hgr/