Learning exceptionality indices for French variable schwa deletion

Aleksei Nazarov (University of Toronto)

An adult phonological grammar must account both for gradient lexicon-wide generalizations and for exceptional words' phonological behavior (Coetzee and Pater 2011). To acquire such a grammar, the language-learning infant must identify the exceptions to each generalization and account for them in the grammar. This exception induction process has previously been computationally modeled in OT with lexically indexed constraints (Kraska-Szlenk 1995, Pater 2000), but these models (e.g., Becker 2009) cannot account for variably applying processes. Moore-Cantwell and Pater (2016) showed that variable processes with exceptions can indeed be learned given the existence of lexically indexed constraints for every word, but they do not address the resulting growth of the number of constraints in the grammar. Nazarov (2018) was the first to show how a limited number of lexically indexed constraints for groups of lexical items (instead of an indexed version of each constraint for each lexical item) can indeed be induced for variable processes. At the same time, this learner did not match the process' rate of application (which is crucial for gradient processes), and the model was not tested on an extended, realistic data corpus. In this paper, I apply a variant of Nazarov's (2018) learner to a corpus of French schwa deletion data (Racine 2008), showing that it can generate appropriate schwa deletion rates in various contexts while also finding exceptions to these rates.

French has a variable schwa deletion process that has both segmental restrictions and lexical exceptions that affect the process' rate of application (Dell 1985). While schwa deletion is generally blocked when it creates certain types of CCC clusters (1a; Racine 2008), it generally applies at a rate above 50% between single consonants, (1b; Racine 2008). However, schwa deletion is dispreferred when it creates a [ɲC] cluster or a word-initial cluster, (1c; Racine 2008).

(1) a. [bøgləmɑ̃, *bøglmɑ̃] b. [bʁɑ̃ʃəmɑ̃ 14% < bʁɑ̃ʃmɑ̃ 86%] c. [bəzwɛ̃ 85% ~ bzwɛ̃ 15%]

Each of these generalizations has exceptions: words where schwa deletion is permitted despite creating a CCC cluster, (2a), words where schwa deletion is (gradiently) blocked from applying between two single consonants, (2b), and words where schwa deletion is preferred despite creating a word-initial cluster, (2c). Examples are taken from Racine (2008).

(2) a. [lɛʃəfʁit 8% < lɛʃfʁit 92%] b. [apəzɑ̃tœʁ, *apzɑ̃tœʁ] c. [ʃəmiz 41% < ʃmiz 58%]

These data were presented to an updated variant of Nazarov's (2018) model of indexed constraint induction, which uses Expectation Driven Learning (EDL; Jarosz 2015) to learn variable OT grammars, but adds machinery that can add to the grammar a series of lexically indexed constraints relevant to the data. The latter can account for the phonological behavior of exceptional words, even if exceptions are not marked as such in the data.

EDL represents within-word variation with probabilistic constraint ranking, similarly to Boersma (1998). Specifically, probabilities are represented over pairwise constraint rankings: P(A >> B), P(A >> C), … An additional mechanism ensures that pairwise rankings sampled from these probabilities always assemble into a logically consistent ranking (Jarosz 2015). Given initial values for these ranking probabilities, the learner successively re-estimates new probabilities given the data and the old probabilities with updates based on Expectation Maximization (Dempster et al. 1977), schematically as in (3). Through sampling of rankings from the grammar, the relative compatibility of every pairwise ranking with a given data point

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 385-386.
New York City, New York, January 3-6, 2019

can be estimated – P(datum$_i$|A>>B) –, and from this, P(A>>B|datum$_i$) can be computed through Bayesian math (Jarosz 2015). From P(A>>B|datum$_i$) values for each individual word, a corpus-wide value P(A>>B|data) is computed, and P(A>>B) in the grammar is updated to that value.

(3) $P_{t+1}(A \gg B) = P(A \gg B|data), computed\ using\ P_t(A \gg B)$

Nazarov's (2018) exception induction proposal uses the fact that the probability of A >> B is computed both given a single data point and given the entire data set. By default, Nazarov (2018) assumes that there are no exceptions, but whenever a word's P(A >> B|datum) is opposite to the entire lexicon's P(A >> B|data), that word is seen as exceptional with respect to A >> B:

(4) $exc(datum_i, A \gg B) \leftrightarrow [P(A \gg B|datum_i) > 0.5 + \theta\ \&\ P(A \gg B|data) < 0.5 - \theta]$

At every iteration of the learner, a single indexed constraint is added or maintained for all words exceptional w.r.t. the pairwise ranking with the greatest difference between exceptions and the lexicon, as in (5). The current model includes Jarosz' (2015) machinery that lets the learner match rates of application, an aspect not included in Nazarov (2018).

(5) $add/maintain\ B_i\ and\ i = \{d|exc(d, A \gg B\}\ such\ that\ A \gg B = \underset{X \gg Y}{\text{argmax}} \sum_{\{z|exc(z,X \gg Y)\}}(|P(X \gg Y|z) - P(X \gg Y|data)|)$

The updated model was tested on data from Racine's (2008) study on French nouns with precisely one word-internal underlying schwa. The training data consisted of the judgments of 12 speakers from France on all 1525 non-compound nouns in the study in their variants with and without schwa, transformed into pseudo-frequencies for each form (since the learner matches pseudo-frequencies for phonological output candidates) by subtracting 1 from the judgments to reflect that '1' was the lowest judgment category, and multiplying by the relevant word's film/book frequency in LEXIQUE 3.30 (New et al. 2001). The following constraints were used:

*OversizedCluster (one violation for each CCC cluster, except C(ə)Cʁ and C(ə)Cl);
*#CC (one violation for every word-initial CC cluster);
*ŋC (one violation for every [ŋC] cluster); *ə; and Max(ə).

The learner was run 10 times on these data, with 30 iterations sans indexed constraints to find lexicon-wide patterns (Prince and Tesar 2004, Jarosz 2006) followed by 10 iterations with indexed constraint induction. Initial probabilities were 0.5 for every constraint pair, and $\theta = 0.1$.

The resulting grammars all had a 95% fit to the training data, with fit defined as matching rates of application: exp[-D_{KL}(predicted||observed)]. Of the 31 words with outlier schwa deletion rates (based on mean schwa deletion rates per phonological context), all 10 runs marked the same 27 words (87%) as exceptions, with an overall 99.7% accuracy in identifying whether a word is an exception. At 8 of 10 runs, the only indexed constraints added to the grammar were *ə$_i$ and Max(ə)$_j$. Thus, a version of Nazarov's (2018) proposal can indeed generate a gradient process's rate of application per context, while finding exceptions to it in a realistic data corpus.

Selected References Moore-Cantwell, C. and J. Pater. 2016. Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints. *Catalan Journal of Linguistics* 15, 53-66. ♦ Nazarov, A. 2018. Learning within- and between-word variation in probabilistic OT grammars. In G. Gallagher, M. Gouskova, and S. Yin (eds.), *Supplemental Proceedings of the 2017 Annual Meeting on Phonology*, LSA. ♦ New, B., Pallier, C., Ferrand, L. et Matos, R. 2001. Une base de données lexicales du français contemporain sur internet: Lexique. *L'Année psychologique* 101, 447-462. ♦ Racine, I. 2008. Les effets de l'effacement du Schwa sur la production et la perception de la parole en français. PhD thesis, Université de Genève.