Incorporating phonological and syntactic factors into sentence probability distributions

A growing body of work indicates that phonological factors can have sentence-level effects, from influencing choices amongst synonymous syntactic alternatives (Shih & Zuraw 2017) to producing outright ineffability (Rice 2007); the effects appear to be widespread and involve multiple constraints (Breiss & Hayes 2019). Importantly, these findings cannot be attributed to general sentence-level preferences for particular phonetic/phonological patterns, because cross-linguistic comparisons reveal that the constraints active at the sentence level in a given language are the same ones that apply word-internally in that language -- for English, Breiss & Hayes find effects for constraints such as *SibilantClash and *Hiatus that are active in the word phonology. This work raises the question of what sort of probabilistic model of grammar could simultaneously express both standard syntactic patterns and cross-word phonological effects. We offer one proposal here.

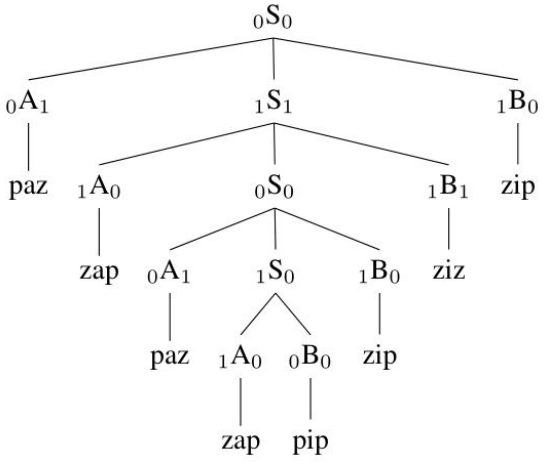
As a simplified artificial example, we consider adding a constraint penalizing adjacent sibilants across word boundaries to the simple context-free grammar (CFG) shown in (1), which generates an $A^n B^n$ pattern. We choose this as a stand-in for the kind of necessarily hierarchical patterns that motivate non-finite-state formalisms for natural language syntax.

(1) S→A S B
S→A B
A→pap | paz | zap | zaz
B→pip | piz | zip | ziz

The sentence in (2), for example, is generated by this grammar, and in addition contains three violations of *SibilantClash (*SC). The challenge is that because the CFG's rules work hierarchically and not linearly, they do not "see" adjacent pairs of words. The key idea that we build on is to locate this string's second *SC violation not at a decision to put the word *zap* after the word *paz*, but rather at the decision to combine the A constituent *paz* with the S constituent *zap pip* as part of applying the S→ASB rule.

(2) paz zap paz zap pip zip ziz zip

To do this, we identify an "annotated tree structure" for any sentence generated by this grammar; this structure for sentence (2) is shown here. Each internal node has binary subscripts on its left and right, indicating whether there is a sibilant at the beginning and/or end (respectively) of the corresponding constituent. The lowest S node, for example, has a 1 subscript on its left to indicate that the left edge of the string *zap pip* is a sibilant, and a 0 subscript on its right to indicate the right edge is not.



In this representation, each *SC violation can be associated with one application of a particular syntactic rule, and can be assessed locally using the information in the subscripts. For example, the *SC violation incurred by forming the constituent *paz zap pip zip* is detectable from the 1s that appear as the right subscript of the daughter $_0A_1$ node and as the left subscript of the daughter $_1S_0$ node. Forming the parent, six-word constituent, from categories $_1A_0$, $_0S_0$ and $_1B_1$ (in that order), incurs no violations. The topmost rule application incurs two violations: one from concatenating a $_0A_1$ with a $_1S_1$, and one from concatenating a $_1S_1$ with a $_1B_0$. (Note that subscripts do not encode violations; they encode phonological information relevant to *detecting* violations.)

In a standard probabilistic CFG, the probability of (2) (leaving aside lexical probabilities) would be (3); to include sensitivity to *SC, a central idea behind our model is to instead use (4).

(3) $\Pr(S \rightarrow ASB) \times \Pr(S \rightarrow ASB) \times \Pr(S \rightarrow ASB) \times \Pr(S \rightarrow AB)$

(4) $\Pr({}_0S_0 \rightarrow {}_0A_1\ {}_1S_1\ {}_1B_0) \times \Pr({}_1S_1 \rightarrow {}_1A_0\ {}_0S_0\ {}_1B_1) \times \Pr({}_0S_0 \rightarrow {}_0A_1\ {}_1S_0\ {}_1B_0) \times \Pr({}_1S_0 \rightarrow {}_1A_0\ {}_0B_0)$

But crucially, the probabilities multiplied together in (4) are not independent parameters associated with primitive rules in a new, "combinatorially exploded" CFG, but rather are determined by the values of other, primitive parameters on the basis of: (i) which underlying CFG rule they realize, and (ii) how many violations of *SC they incur. Thus, (4) reduces to (5).

(5) $\Pr(S \rightarrow ASB,\ 2\ \text{viols.}) \times \Pr(S \rightarrow ASB,\ 0\ \text{viols.}) \times \Pr(S \rightarrow ASB,\ 1\ \text{viols.}) \times \Pr(S \rightarrow AB,\ 0\ \text{viols.})$

Specifically, we use maxent/log-linear models that share parameters to define each of the local multinomial distributions over rewrite rules that are primitive in a standard probabilistic CFG (Berg-Kirkpatrick et al. 2010). We include an indicator feature for each original CFG rule, and a feature for each phonological constraint: here, an indicator feature for S→ASB, an indicator feature for S→AB, and a feature for *SC violations. The probability of a particular "enriched rule" such as ${}_0S_0 \rightarrow {}_0A_1\ {}_1S_1\ {}_1B_0$ is therefore determined by tradeoffs between independently weighted factors, some syntactic (is this a use of the ASB rule?) and some phonological (how many times does this violate *SC?) -- like candidate probabilities in MaxEnt OT (Goldwater & Johnson 2003).

Each weight inferred from a training corpus therefore affects a wide range of "enriched rules", producing appropriate generalizations. The table reports a toy example. The preference for avoiding *SC violations, although it only surfaces amongst the S→ASB rules in the training data, is carried over to the S→AB rules where it was not observed. Similarly, although one-violation and two-violation instances of S→ASB were equally frequent in the training data, their rarity relative to the zero-violation option leads to a lower probability for two-violation instances.

	Training frequency	Learned probability
S→AB, 0 violations	10%	0.15
S→AB, 1 violation	10%	0.05
S→ASB, 0 violations	60%	0.53
S→ASB, 1 violation	10%	0.20
S→ASB, 2 violations	10%	0.07

There are no barriers to scaling this method up: we can add more phonological constraints (by replacing singleton subscripts with vectors), and can apply the approach to large-coverage CFGs, or to other formalisms such as Multiple Context Free Grammars (Seki et al. 1991, Stabler 2011) to incorporate movement phenomena. The approach allows independently motivated phonological and syntactic preferences to compete, without invoking "mixed" constraints that could, e.g., require movement of constituents beginning with a bilabial (Zwicky & Pullum 1986). This opens up the possibility of more rigorously testing the degree to which this kind of constrained interaction is empirically justified -- for example, building on the large-scale tests of Breiss & Hayes 2019 by replacing their bigram-based model of word combination with a more realistic syntax.

Berg-Kirkpatrick et al. 2010, "Painless unsupervised learning with features", NAACL 2010. **Breiss & Hayes 2019**, "Phonological markedness effects in sentence formation", lingbuzz/004487. **Rice 2007**, "Gaps and repairs at the phonology-morphology interface", *J. Linguistics*. **Goldwater & Johnson 2003**, "Learning OT constraint rankings using a maximum entropy model", *Stockholm workshop on variation within OT*. **Seki et al. 1991**, "On multiple context-free grammars", *Theoretical Computer Science*. **Shi & Zuraw 2017**, "Phonological conditions on variable adjective and noun word order in Tagalog", *Language*. **Stabler 2011**, "Computational perspectives on minimalism", *Oxford Handbook of Minimalism*. **Zwicky & Pullum 1986**, "The principle of phonology-free syntax", *OSU Working Papers in Linguistics*.