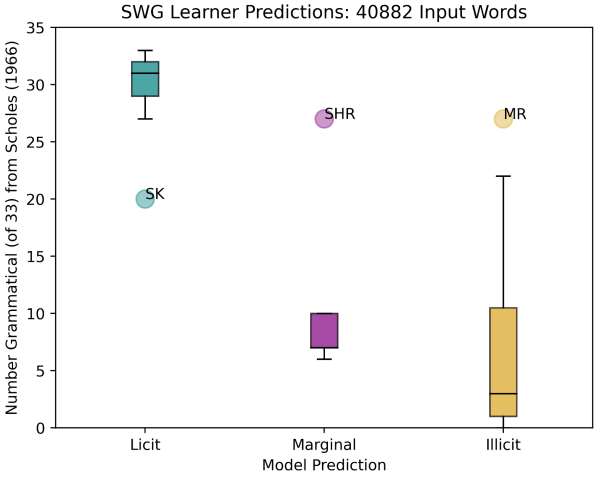
Marginal Sequences are Licit but Unproductive

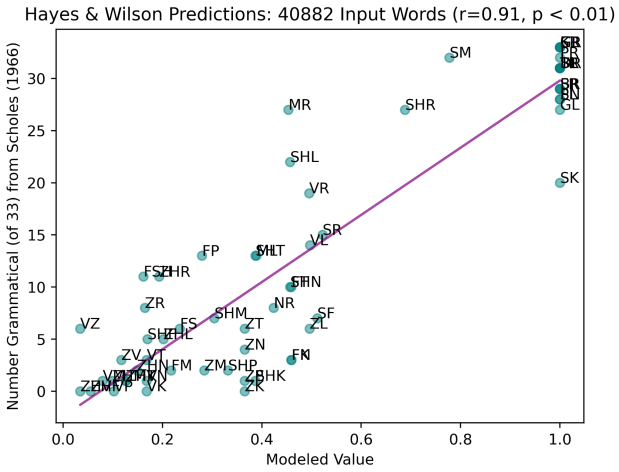
Introduction: Theories of phonotactics (e.g. 1, 2) generally assume that attested sequences are licit and unattested ones are illicit. However, marginal forms (e.g. English ?[#sf]) do not fit neatly into this dichotomy: while attested, they receive low ratings on wordlikeness tasks (e.g. 3). Some approaches (e.g. 2) view marginal forms as an exceptional (i.e. not unattested) subclass of illicit forms. However, marginal forms pattern more like licit ones in terms of repairs in borrowings (e.g. ?[#sf] is not repaired in "sphere" or "sphinx," patterning like [#sp] in "spaghetti" rather than *[#pn] in "pneumonia") and production and perception errors (4, 5) suggesting that marginal forms are actually a subclass of *licit* forms. We argue for a theory of phonotactics in which attested sequences are divided into *productive*, licit ones and *unproductive*, marginal ones. We present a syllable-based computational model that classifies forms as marginal, licit, or illicit, and show that this model matches well with human judgments while accounting for the patterning of marginal forms with licit ones.

Model: We propose the *Sequence-wise Generalization* learner (SWG), a computational model for learning phonotactics. SWG is motivated by evidence that phonotactic knowledge is represented over syllables (6) and the observation that licit onsets/codas generally occur with a wider range of nuclei than marginal ones. For example, [#sp] is licit, occurring before 15/15 possible nuclei in our full training data (below), but ?[#sf] is marginal, occurring before only 2/15. SWG formalizes this observation via the **Tolerance/Sufficiency Principle** (TSP, 7): a sequence attested with M distinct nuclei is licit iff $N\text{-}M \leq N/\ln N$, where N is the number of possible nuclei in the language. SWG is also motivated by evidence that early representations are underspecified (8) and learns phonotactics as increasingly-specific sequences of feature sets. At each step of learning, SWG intersects all sequences of feature-sets in the current input to yield an underspecified sequence S. If sufficiently many sequences of segments matching S are attested in the input and licit, SWG adds S to the set of licit sequences in its grammar. Otherwise, it divides the input based on the most frequent feature at the position with the greatest difference between N and M, recursively learning from each resulting set. If no generalization is made and no more features are available to subdivide on, remaining input sequences are memorized as marginal.

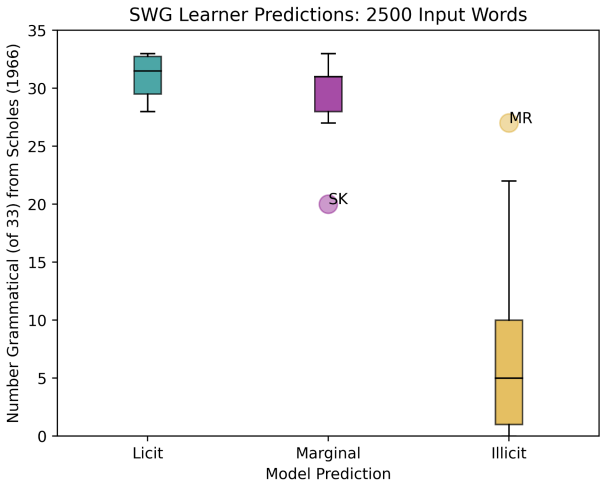
Results: We evaluate SWG on English complex onsets, comparing the grammaticality judgments of 7th graders from (3) to SWG's predictions. To simulate a plausible vocabulary, we intersect CELEX (9) with the CMU pronouncing dictionary (10), syllabified using (11), yielding 40,882 words. For each onset in (3), we compare the model's categorical prediction (licit/marginal/illicit) to the number of students (of 33) who rated it as licit. When trained on all words – an approximation of a 7th grader's vocabulary (12) – SWG matches well with human judgments (Figure 1, Pearson's $r(60) = 0.86$, $p < 0.0001$) and is comparable in performance to the model of (2) (Figure 2, average correlation across 5 runs: $r(60) = 0.91$, $p < 0.0001$). SWG also predicts a significant difference between licit and marginal forms (Welch's $t(4.3) = 4.6$, $p < 0.01$). When trained on only 2,500 forms – a reasonable approximation of vocabulary before 4;0 (13) – SWG still correlates well with human judgments (Figure 3, Pearson's $r(60) = 0.83$, $p < 0.0001$), and is again comparable in performance to the model of (2) (Figure 4, average correlation across 5 runs: $r(60) = 0.83$, $p < 0.0001$). However, SWG does not predict a significant difference between licit and marginal forms (Welch's $t(15) = 1.25$, $p = 0.23$), but classifies most complex onsets as marginal. SWG thus predicts that phonotactic acquisition proceeds as an early

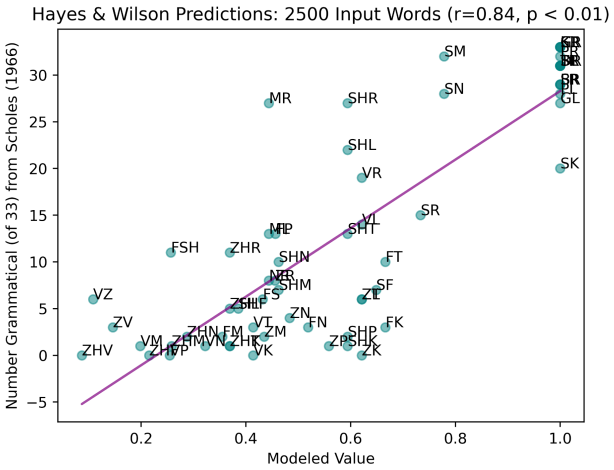
stage of memorization followed by a later stage of generalization; this can be tested empirically in future work. In sum, SWG provides a syllable-based, learning-based account of marginal vs. licit sequences as a difference in productivity. In contrast to previous models such as (2), in which marginal forms are conceptualized as a subclass of illicit forms, SWG accounts for the patterning of marginal forms like licit ones in borrowings and production and perception errors, while still matching well with human judgments.





Figures 1 & 2: Output of SWG (left) and the best Hayes & Wilson run (right) on full train





Figures 3 & 4: Output of SWG (left) and the best Hayes & Wilson run (right) on 2500 words

References: **[1]** Chandlee, Eyraud, Heinz, Jardine & Rawski 2019. *Proceedings of Mathematics of Language.* **[2]** Hayes & Wilson 2008. *Linguistic Inquiry.* **[3]** Scholes 1966. *De Gruyter Mouton.* **[4]** Davidson 2006. *Journal of Phonetics*. **[5]** Gorman 2013. *PhD Dissertation, University of Pennsylvania.* **[6]** Kabak & Idsardi 2007. *Language and Speech.* **[7]** Yang 2016. *MIT Press*. **[8]** Werker, Fennell, Corcoran, & Stager 2002. *Infancy.* **[9]** Baayen, Piepenbrock, & Gulikers 1996. *The CELEX Lexical Database.* **[10]** Weide 1998. *The CMU Pronouncing Dictionary.* **[11]** https://github.com/kylebgorman/syllabify **[12]** Nagy & Anderson 1984. *Reading Research Quarterly.* **[13]** Fenson, Dale, Reznick, Bates, Thal, Pethick, Tomasello, Mervis, & Stiles 1994. *Monographs of the Society for Research in Child Development.*