Phonotactic learning in the presence of exceptions with a categorical approach

Lexicalized exceptions are a major source of noise in phonological acquisition. In a positive-evidence-only setting, it is common to cope with exceptions with *indirect negative evidence* from distributional information (Clark & Lappin 2010). Most distribution-sensitive models assume a probabilistic grammar that evaluates the grammaticality of words by their predicted likelihood (Hayes & Wilson 2008). However, a probabilistic grammar conflates all words into the same spectrum of probability and grammaticality. As a result, short attested exceptions become more 'grammatical' than longer grammatical words with lower probabilities (Daland 2015). This is problematic because it blurs the boundary between exceptions and grammatical words. In this paper, I spell out a formal language-theoretic ('memory-based' per Wilson & Gallagher 2018) algorithm that learns a categorical grammar in the presence of exceptions *w.r.t.* a (Tier-based) Strictly Local-2 (SL_2; Rogers & Pullum 2011) hypothesis space and O/E criterion (Frisch et al. 2004). I argue that this approach is at least as good as, and in one case appears to be superior to the "Probabilistic grammar + Probabilistic inference" approaches (Hayes & Wilson 2008) in handling exceptions.

1. Proposal. Consider a toy example of CV structure: for a given inventory $\{C, V\}$, the hypothesis space $\textsc{Con} = \{^*VV, ^*VC, ^*CC, ^*CV\}$ includes all possible forbidden adjacent sequences of length two ('2-factors'). The learning objective is to acquire the target grammar $G = \{^*CC, ^*VV\}$ from a sample $S = [VC, CVC, VCV, \mathbf{VV}]$ with an exception **VV** of low frequency. Two halves in Table (a) show the computation **before** and **after** the learning procedure: (1) initializing a full hypothesis grammar H that is identical to $\textsc{Con}$; (2) computing statistics (O/E); (3) removing a constraint from H if its statistical criterion O/E is larger or equal to 1 ('overrepresented'). Previous works applied O/E as heuristics in probabilistic models, in which exceptions receive nonzero probabilities (Gouskova & Gallagher 2020). For a constraint $^*\sigma_1\sigma_2$, the **observed** frequency of violations (O) is the sum of $\sigma_1\sigma_2$ occurrences in S. The **expected** frequency of violations (E) is $\frac{N(\sigma_1)*N(\sigma_2)}{N(\sigma_i\sigma_j)}$. $N(\sigma_1)$ and $N(\sigma_2)$ respectively correspond to the frequency of σ_1 and σ_2 occurrences in the first and second position of 2-factors in S. $N(\sigma_i\sigma_j)$ denotes the frequency of all 2-factors in S. The learned grammar H converges to $\{^*CC, ^*VV\} = G$, as indicated by grey cells in (a). The current proposal utilizes distributional information through O/E and embraces categorical grammaticality (Chomsky & Halle 1965)—the attested exception **VV** is ungrammatical due to its underrepresented frequency ($O/E = 0.5$), and categorically penalized by the learned grammar.

2. Turkish. I applied the proposed algorithm to the Turkish Electronic Living Lexicon (Inkelas et al. 2009; Gouskova & Stanton 2021; TELL), which consists of $\approx 66,000$ roots and elicited derived forms. Two productive constraints trigger progressive harmony across morpheme boundaries: (1) a vowel cannot follow another vowel with a different [back] value; (2) a high vowel cannot follow another vowel with a different [round] value. Turkish vowel harmony is known for its exceptions, especially *labial attraction* where $aC_{[+\text{lab}]}u$ is produced due to the intervocalic labial consonant, e.g. *sabur* 'patient' (Lees 1966). This pattern, however, is **not internalized** by native speakers (Zimmer 1969); in other words, they are attested but ungrammatical (Gorman 2013). The computed O/E of the **learned grammar** of vowel agreement patterns are highlighted in grey cells in Table (b). The learned grammar successfully generalizes vowel harmony patterns in Turkish, and categorically penalizes the exception $a \ldots u$ in labial attraction. There are several marginal cases, **boldfaced** in (b), caused by underrepre-

$^*\sigma_i\sigma_j$	O	E	O/E
*VV	0	0	0
*VC	0	0	0
*CC	0	0	0
*CV	0	0	0
*VV	1	2	0.5
*VC	3	2	1.5
*CC	0	1	0
*CV	2	1	2

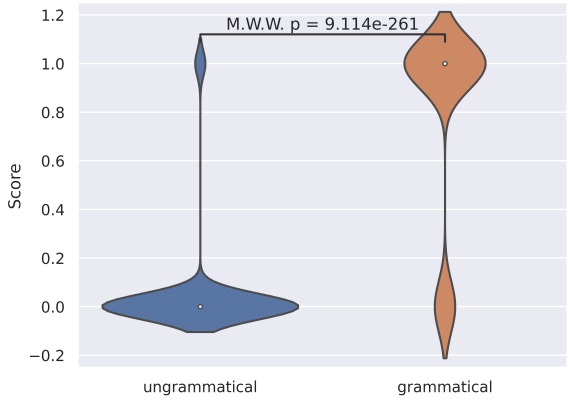
(a) Toy CV example

$\sigma_i \downarrow \sigma_j \rightarrow$	i	e	y	ø	ɯ	a	u	o
i	2.15	1.25	0.21	**0.92**	0.16	0.65	0.10	**1.26**
e	2.04	1.71	0.44	**0.76**	0.11	0.46	0.24	0.75
y	0.20	2.32	9.77	**0.97**	0.08	0.46	0.26	0.22
ø	0.05	2.59	11.21	2.03	0.04	0.20	0.09	0.24
ɯ	0.07	0.20	0.03	0.86	3.36	1.18	0.09	**0.49**
a	0.40	0.48	0.21	**1.32**	2.08	1.49	0.43	1.18
u	0.18	0.38	0.33	0.73	0.07	1.49	6.33	**0.47**
o	0.36	0.55	0.35	0.84	0.06	1.39	4.93	2.51

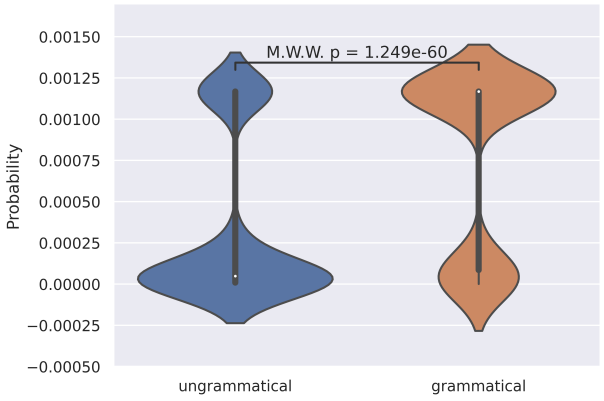
(b) Learning Turkish vowel harmony patterns

sented grammatical words and overrepresented exceptions, which exhibit the limitation of a stringent O/E criterion. This issue might be resolved by incorporating more sophisticated criteria, such as *gain* (Gallagher & Gouskova 2020; *cf.* Stanton & Stanton 2022).

3. Comparison. The current proposal and Hayes & Wilson (2008)'s MaxEnt learner are trained on ≈ 66,000 words in TELL and evaluated on 2,000 manually labelled nonce words in Turkish, following Gouskova & Gallagher (2020). I mapped labelled and predicted categorical judgements of grammaticality to scores 0 and 1. Spearman's correlation between labelled grammaticality and predicted distribution in MaxEnt learner is $\rho = 0.366$, which is lower than 0.771 in my proposal. Figures (c) and (d) show the result of a Mann-Whitney-Wilcoxon (M.W.W.) test that measures whether two distributions are significantly distinguishable. Although both learners significantly distinguish grammatical and ungrammatical words, the MaxEnt learner results in more errors of classification because it assigns nonzero probabilities to exceptions and consequently penalizes productive patterns in grammatical words.



(c) The current proposal



(d) MaxEnt over [+syllabic] tier

Moreover, probabilistic grammar predicts gradient productivity, in which exceptions might surface and become productive (Moore-Cantwell & Pater 2016). However, given the counterexample of the uninternalized labial attraction, it is disputable whether the exceptional patterns in Turkish are truly productive with current experimental evidence (Gorman 2013).

4. Conclusion. If a learner concludes that everything, including noise, is grammatical, then nothing is learned. The 'categorical grammar + statistical criterion' approach provides an explicit demarcation of exceptions and grammatical words, eliminating the need for a special status of exceptions in a probabilistic grammar. This proposal provides a compelling alternative to the long-standing problem of phonotactic learning in the presence of exceptions.