Malagasy OCP targets a single affix: implications for morphosyntactic generalization in learning

Investigators have uncovered evidence for phonological LEARNING BIASES: biases inherent in learners that favor certain language phonologies over others (Wilson 2006, Finley 2012, Moreton & Pater 2012, Hayes & White 2013, White 2014, McMullin & Hansson 2014, a.o.). How strong are these biases? In particular, to what extent can a learning bias be defied in language? These questions bear directly on the theory of phonological learning, as they address the limits of learner capability.

A growing family of findings suggests that learners tend to favor phonological constraints that are morphosyntactically general—i.e., that are obeyed by at least several morphemes, or in multiple or all grammatical contexts. That phonological alternations are typically corroborated by the phonotactic constraints of a given language was observed as early as Chomsky & Halle (1968) and Kenstowicz & Kisseberth (1977), but the generalizing tendency just mentioned has also been observed in a number of recent corpus studies: Martin (2007, 2011), Zuraw (2015), Breiss & Hayes (2018), and Shih & Zuraw (in press) observe cases of grammatical “leaking”, in which strong phonotactic restrictions tend to manifest across word boundaries or compound boundaries, or affect the choice between grammatical constructions; Chong (2016) found that certain famous apparent derived environment effects are just that—only apparent. Generalization effects were also borne out in artificial language learning experiments: Myers & Padgett (2014) found that participants generalize a phrase-final devoicing pattern to the word-final domain without exposure to unambiguous evidence; Chong (2017) found that participants more readily learned a suffixal harmony alternation when they were exposed to higher
rates of root harmony, corroborating proposals that phonotactic generalizations assist in acquiring alternations (Tesar & Prince 2003, Hayes 2004, Jarosz 2006, a.o.).

This squib complicates our current understanding of learners’ tendency to posit morphosyntactically general constraints. Malagasy displays backness dissimilation, an alternation that has persisted across multiple generations that sends a back vowel to front in the presence of a nearby back vowel. The process applies very consistently to the passive imperative suffix, –u, and displays blocking behavior typical of dissimilation, suggesting the working of an OCP constraint. But –u is the only affix in the language that undergoes dissimilation, and is the only suffix even eligible to undergo it. Moreover, stems in the lexicon show no preference for dissimilation whatsoever; in fact, they display a modest but highly significant opposing preference for harmony. This suggests that Malagasy learners induce a morphologically specific OCP constraint—specific either to –u alone or to the suffix domain as a whole—without the need for a corroborating phonotactic trend. These findings suggest that no degree of morphosyntactic generality is a necessary condition for learning. Though learners might be biased towards acquiring grammatically general constraints, the Malagasy system suggests that they are capable of overriding this bias completely. I present this system below, and discuss the problems it poses for a theory in which learners favor grammatically general constraints.

1. **Backness dissimilation applying to the passive imperative suffix**

Unless otherwise specified, the data below come from the Malagasy Dictionary and Encyclopedia of Madagascar (hereafter MDEM; malagasyword.org; de la Beaujardière 2004), an annotated online corpus containing ~92,000 Malagasy words. The Malagasy
vowel inventory is composed of [i e a u] (Parker 1883, de la Beaujardière 2004). There are four suffixes: the passive suffixes –ina and –ana, the active imperative suffix –a, and the passive imperative suffix –u (Parker 1883, Richardson 1885).

The passive imperative suffix conditionally undergoes backness dissimilation (Parker 1883, AUTHOR 2015): underlying –u (1a-b) surfaces as –i after stems containing u (2a-d) unless a front vowel intervenes (3a-b). The alternation conforms to patterns driven by the Obligatory Contour Principle (Leben 1973, Goldsmith 1976, et seq).

| Underlying –u | (1a) /bata+u/ [bata-u] lift-PASS.IMP | (1b) /sava+u/ [sava-u] inspect-PASS.IMP |
| Backness dissimilation | (2a) /babu+u/ [babu-i] plunder-PASS.IMP | (2b) /tuv+u/ [tuv-i] fulfill-PASS.IMP |
| | (2c) /suav+u/ [suav-i] bless-PASS.IMP | (2d) /u°dan+u/ [u°dan-i] bolster-PASS.IMP |
| Blocking by front vowels | (3a) /turi+u/ [turi-u] preach-PASS.IMP | (3b) /fules+u/ [fules-u] thread-PASS.IMP |

3,675 words in MDEM with the passive imperative suffix were extracted. The counts in Table 1 show that dissimilation is triggered by the presence of stem-internal u, applies regularly when the trigger is local and semi-regularly across a, and is regularly blocked by front vowels.

<table>
<thead>
<tr>
<th>Context (ignoring consonants)</th>
<th>–u</th>
<th>–i</th>
<th>Dissim. rate</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>No trigger</td>
<td>1877</td>
<td>7</td>
<td>0.0%</td>
<td>bata-u</td>
</tr>
<tr>
<td>Adjacent trigger</td>
<td>4</td>
<td>989</td>
<td>99.6%</td>
<td>babu-i</td>
</tr>
<tr>
<td>Intervening a</td>
<td>196</td>
<td>201</td>
<td>50.9%</td>
<td>tuda-i</td>
</tr>
<tr>
<td>Intervening front vowel</td>
<td>399</td>
<td>2</td>
<td>0.4%</td>
<td>turi-u</td>
</tr>
</tbody>
</table>

Table 1: Counts for Malagasy backness dissimilation
Multiple lines of evidence suggest that Malagasy speakers acquire this alternation. Dissimilation is observed across at least two generations: it was reported as early as Parker (1883), and evidence for it appears in dictionaries since then (e.g., Abinal & Malzac 1888, Rajemisa 1985, de la Beaujardière 2004). Dissimilation and its blocking can be observed even when –u comes after loaned stems, as in (4a-d) below. The stems given below can be found in the World Loanword Database (wold.clld.org; Adelaar 2009), except /matsu/, which is marked as a loan in MDEM.

\[
\begin{align*}
\text{Dissimilation} & \quad (4a) \quad /\text{ma}\tilde{s}u+u/ \quad [\text{ma}\tilde{s}u-i] \quad \text{march-PASS.IMP} \quad \text{English loan} \\
& \quad (4b) \quad /\text{kiraru}+u/ \quad [\text{kiraru}-i] \quad \text{shoe-PASS.IMP} \quad \text{Bantu loan} \\
& \quad (4c) \quad /\text{kuhu}ku\text{hu}+u/ \quad [\text{kuhu}ku\text{hu}-i] \quad \text{cluck-PASS.IMP} \quad \text{Bantu loan} \\
\text{Blocking} & \quad (4d) \quad /\text{burus}i+u/ \quad [\text{burus}i-u] \quad \text{brush-PASS.IMP} \quad \text{French loan}
\end{align*}
\]

Remarkably, the passive imperative suffix is the only affix to undergo dissimilation, and, assuming the process sends back vowels to front but not vice versa, is the only suffix even eligible to undergo it (being the only one to contain u). Even if we assume that dissimilation sends back vowels to front and vice versa, it is still not displayed by any other affix, according to an MDEM search—see the Appendix for details. If there were other evidence for a dissipilatory tendency in the grammar, we would expect to find it in phonotactics. We now turn to a corpus study of roots to assess whether this is the case.

2. A backness harmony trend in Malagasy stem phonotactics

Surprisingly, roots display a modest but highly significant tendency toward backness harmony. MDEM gives numerous harmonic roots:

(5) \begin{align*}
\text{kiri} & \quad \text{‘small hole’} \\
\text{lufu} & \quad \text{‘persistence’} \\
\text{gegi} & \quad \text{‘indiscreet’} \\
\text{sarut’u} & \quad \text{‘cape’} \\
\text{tevika} & \quad \text{‘spasm’} \\
\text{vulu} & \quad \text{‘color’} \\
\text{uzuna} & \quad \text{‘curve’} \\
\text{išindri} & \quad \text{‘compression’} \\
\text{duku} & \quad \text{‘identity’}
\end{align*}
Counts of tier-adjacent pairs involving only front or back vowels (i, e, and u) were enumerated across 4,514 roots that were extracted from MDEM. The counts reveal no preference for disharmonic sequences in roots, as Table 2 reveals below. Note that the majority of roots in the corpus are classified as nouns (2,737), adjectives (729), or adverbs (733); verbs are derived through affixation (cf. Keenan & Polinsky 1998).¹

<table>
<thead>
<tr>
<th></th>
<th># harmonic VCV seq.s</th>
<th># disharmonic VCV seq.s</th>
<th># harmonic VCaCV seq.s</th>
<th># disharmonic VCaCV seq.s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within noun roots</td>
<td>786</td>
<td>602</td>
<td>35</td>
<td>44</td>
</tr>
<tr>
<td>Within adj. roots</td>
<td>185</td>
<td>183</td>
<td>27</td>
<td>11</td>
</tr>
<tr>
<td>Within adv. roots</td>
<td>312</td>
<td>188</td>
<td>109</td>
<td>49</td>
</tr>
<tr>
<td>Within interj., conj., prep. roots</td>
<td>96</td>
<td>41</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1379</td>
<td>1014</td>
<td>205</td>
<td>118</td>
</tr>
</tbody>
</table>

Table 2: Raw counts of (dis/)harmonic sequences in roots

There are around 350 more local harmonic sequences than local disharmonic sequences, and around 100 more nonlocal harmonic sequences than nonlocal disharmonic sequences. This makes backness dissimilation highly morphologically specific: it requires reference to the suffix domain or to the passive imperative suffix in particular, and lacks a counterpart generalization in stem phonotactics.

In fact, as we will see, Malagasy displays a significant tendency toward backness harmony in roots—these counts are unlikely to have arisen by chance alone. The observed rates of local and nonlocal harmony are 1379/(1379 + 1014) = 57.3% and

¹ Some words displaying reduplication (cf. Lin 2005) were classified as roots in the corpus; in these cases, only the root involved in reduplication contributed to the counts, rather than the reduplicated stem as a whole. A conference reviewer points out that there could exist productive pseudoreduplication, with the first syllable being a copy of the second, potentially inflating the harmony rate. The corpus revealed that only 115 of the 4,514 roots have matching first and second syllables, with only 64 beginning with a front or back vowel ([didit’a] = ‘twisting’, [vuvuka] = ‘dust’). It is not at all obvious that the language possesses pseudoreduplication, considering how low the count is here.
205/(205 + 108) = 63.5%, respectively. We can calculate the expected local harmony rate given the frequencies of front and back vowels by extracting from roots all $V_1V_2$ sequences in which each vowel belongs to [i e u], and calculating \[p(V_1 = u) \times p(V_2 = u)\] + \[p(V_1 = i \text{ or } e) \times p(V_2 = i \text{ or } e)\], where e.g. $p(V_1 = u)$ is the number of instances of $u$ in $V_1$ position divided by the number of instances of $i$, $e$, and $u$ in $V_1$ position. The expected rate of nonlocal harmony is computed analogously over $V_1aV_2$ sequences. Doing this, we obtain 51.6% and 57.7% as expected rates of local and nonlocal harmony. Comparing the observed and expected rates, we find that observed rates (local: 57.3%; nonlocal: 63.5%) are higher than expected (local: 51.6%; nonlocal: 57.7%). To determine whether harmonic sequences occur significantly more than chance would predict, we can run a Monte Carlo simulation (Kessler 2001). To run a simulation for local vowel sequences, we gather pairs of tier-adjacent vowels belonging to [i e u], shuffle the second vowels of each pair and randomly concatenate each of them to a first vowel, calculate the new harmony rate, and then repeat 10,000 times. The simulation for nonlocal sequences ($V_1aV_2$) can be computed analogously. Figures 1a-b below show histograms of (non/)local harmony rate frequencies after the 10,000 trials.

Figures 1a & 1b: distribution of non/local harmony rates yielded by Monte Carlo trials
For local harmony, the observed rate of 57.3% is greater than any rate yielded by 10,000 trials, and is thus significantly greater than chance would predict (est. $p < \frac{1}{10000} = 0.0001$). For nonlocal harmony, the observed rate of 63.5% is greater than 9,834 of the trials, and is thus significantly above chance as well (est. $p = \frac{10000-9834}{10000} = 0.008$). The results suggest that overrepresentation is not coincidental, but rather reflects a backness harmony preference in phonotactics. Note that there exists some evidence of a harmony alternation as well. The –in––un–infix is used to create passive verbs ([sava] = clear, [s-in-ava] = clear-PASS; [fidi] = choice, [f-in-idi] = choice-PASS). MDEM gives 288 words with –in– and 14 with –un–. –in– can surface before any vowel, and in particular surfaces before $u$ in 56 forms. But in the 14 forms with –un–, the following vowel is always $u$ ([buri] = round; [b-un-uri] = round-PASS), suggesting that –un– is selected to satisfy a (weak) harmony drive. That an infix can harmonize while another suffix dissimilates is reminiscent of Yucatec Maya, in which two suffixes harmonize for backness and height, but another dissimilates for backness, and yet another for backness and height (Blair 1964; see Krämer 2001 for an account). Altogether, these cases suggest that contradictory markedness preferences can target different morphemes or domains.

To summarize, Malagasy backness dissimilation applies consistently to the passive imperative suffix and displays blocking behavior typical of OCP. Roots, however, show no dissimilatory tendency, but rather a modest but highly significant harmony preference.

3. Discussion and potential analytical directions

The Malagasy system provides evidence that the learner can counteract the tendency to favor morphosyntactically general constraints. This finding patterns with other
instances of learning bias defiance, in which systems that have been suggested to be disfavored by learners occasionally arise in the world’s languages and persist across generations, providing evidence that they can be apprehended to some extent (Hayes, Zuraw et al. 2009; Hayes & White 2015; Merrill 2015; Beguš & Nazarov 2017).

The Malagasy system complicates the picture of how a morphosyntactic generality bias in phonological learning should be modeled. Martin (2011) finds that phonotactic constraints can “leak” into the cross-boundary domain: in Navajo sibilant harmony and English geminate avoidance, a categorical generalization within roots is mirrored by a statistical tendency across compound boundaries. To account for this, Martin introduces a Gaussian smoothing term into a MaxEnt learning system so that when the learner weights positively a structure-specific constraint (e.g., applying only stem-internally), it gives small positive weight to an analogous domain-general constraint, leading over time to a grammar with the morphosyntactic generality property. A model in which the usage of a structure-specific constraint implies the usage of an analogous structure-insensitive constraint cannot be applied to Malagasy, at least without further elaboration.

Two potential solutions to the problem are entertained. The first is to say that while any particular affix is allowed to depart from typical phonological behavior in a language, whole domains must overall respect the generality property, at least to a degree. We can say that OCP targets the passive imperative suffix in Malagasy, rather than the entire suffix domain, and so no generalizing tendency should arise. Although this would be a possible approach, we cannot be sure that OCP targets –u rather than the entire suffix domain: it could be that OCP in Malagasy is triggered only by back vowels and is
indexed to the suffix domain, and thus the one suffix with a back vowel, –u, undergoes dissimilation (see discussion at the end of Section 1). Nonetheless, corpus studies undertaken by Chong (2017) support indexing OCP to –u rather than to its domain, as they discount claims of the existence of certain derived environment effects—that is, domain-level mismatches: though prior investigators show that palatalization in Korean (Kiparsky 1973, 1993; Iverson & Wheeler 1988) and velar deletion in Turkish (Lewis 1967, Sezer 1981) avoid sound sequences that are found in some of the languages’ roots, Chong shows that such roots are underattested in these languages. These languages therefore still display the morphosyntactic generality property, at least for the most part. That being said, it may be that the generalizing tendency is not universal even for domains: Finnish shows no tendency against [ti] sequences in roots, but three suffixes regularly undergo assimilation, and one suffix optionally assimilates, to avoid [t+i] (esp. Chong 2017, Anttila 2006; cf. Kiparsky 1973, 1993; Karlsson 1983). The Finnish system suggests that even domains can, to some extent, mismatch overall.

Another possible solution is to say that a generalizing bias even applies in the Malagasy case, but that Malagasy learners make use of a harmony constraint that counteracts leaking of the dissimilatory drive into phonotactics. One can imagine that a learner with a generalizing bias, upon encountering the Malagasy system, would invoke a morphologically specific OCP constraint, and then “smooth” over the grammar with a general OCP constraint, so that the dissimilatory drive leaks into stems. This alone could not account for the Malagasy system, since no dissimilatory tendency is observed in phonotactics. Thus, to correct for this, the learner could weight positively a harmony
constraint so that the phonotactic dissimilatory tendency is cancelled or overridden (see AUTHOR 2018 for a MaxEnt model involving this).

Some evidence indeed suggests that learners can make use of constraints driving dissimilation in some morphemes or domains but harmony in others: after all, Malagasy displays consistent dissimilation to the passive imperative suffix, but a harmony tendency in phonotactics; in addition, backness dissimilation and harmony constraints seem to condition allomorphy in different suffixes in Yucatec Maya (Blair 1964, Krämer 2001). One might wonder, then, why contradictory-preferences systems are so typologically infrequent. Perhaps they are tied to \textit{backness restriction} in particular. The cases of leaking found in Martin (2011) involve sibilant harmony and geminates; considering that grammars \textit{preferring} disharmonic sibilants or geminates are rare or unattested, we might imagine that learners would not entertain such preferences as hypotheses about different grammatical contexts. As a result, sibilant harmony or geminate avoidance found in one grammatical context would leak into another. But backness harmony \textit{and} dissimilation are observed crosslinguistically (Parker 1883, Esztergár 1971, Campbell 1977, Clements \& Sezer 1982, Itô 1984, Harrison 1999, \textit{a.o.}), and so it may be that the learner can entertain constraints driving both backness harmony and dissimilation in hypotheses about these different contexts. Learners might spread the effect of one of these constraints across contexts (e.g., dissimilation), but counteract the effect using the natural opposing constraint (harmony). It could be that generalization effects are only defied in cases where there exists crosslinguistic evidence for the working of two opposing constraints,
as in backness dissimilation and harmony. Where there does not, languages requiring restrictions specific to grammatical context may be relatively prone to being generalized.

How might the Malagasy system have arisen if the generalizing bias is true? Here the picture is unclear, but we can speculate: the passive imperative suffix may have been adopted late in the language’s development, with dissimilation arising to distinguish the suffix boundary—a drive for recoverability that would directly conflict with the generalizing bias. Or perhaps dissimilation began as a constraint against $u+u$ sequences, mirroring a ban on pairs of directly adjacent $u$’s in phonotactics, but was somehow generalized to $u…+u$ sequences. This is a topic here left for further research.

4. Conclusion

Several findings now suggest that learners tend to favor morphosyntactically general phonological constraints. This squib argues that this bias, if it exists, can be overridden. Malagasy backness dissimilation applies very consistently to the passive imperative suffix $–u$, and displays blocking behavior typical of OCP. But $–u$ is the only affix in the grammar that undergoes it, and is the only suffix even eligible to undergo it. Stems, on the other hand, display a modest but significant harmony trend. This suggests that Malagasy learners induce a morphologically specific OCP constraint—specific either to $–u$ or to the suffix domain as a whole—without the need for a corroborating phonotactic trend. These findings suggest that no degree of morphosyntactic generality is a necessary condition for learning. Though learners might favor grammatically general constraints, the Malagasy system suggests that they are capable of overriding this bias completely.
Appendix

Given below are all affixes in MDEM occurring with at least 20 stems and that can place a front/back vowel tier-adjacent to a front/back root vowel. None of them alternate based on the root vowel (see http://malagasyword.org/bins/derivLists?form#longScroll), except for –in–/–un–, which displays some evidence of a harmony alternation (see Section 3).

<table>
<thead>
<tr>
<th>Pref.</th>
<th># forms w/ pref.</th>
<th>Circumf.</th>
<th># forms w/ circumf.</th>
<th>Inf.</th>
<th># forms w/ inf.</th>
<th>Suff.</th>
<th># forms w/ suff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>fi-</td>
<td>2618</td>
<td>fi-…-ana</td>
<td>2144</td>
<td>-in/-un-</td>
<td>288+14</td>
<td>-ina/-na²</td>
<td>1700+32</td>
</tr>
<tr>
<td>ki₃</td>
<td>78</td>
<td>i-…-ana</td>
<td>1991</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ml-</td>
<td>4312</td>
<td>aᵐpi-…-ina</td>
<td>renders X into relative verb</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mᵖi-</td>
<td>1975</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tsi-</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ku-</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>faᵐpi-</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maᵐpi-</td>
<td>693</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mᵖaᵐpi-</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: different frequently occurring affixes and their counts

² The counts of the –na allomorph might be inaccurate, as it also serves as the allomorph to another passive suffix, –ana (Richardson 1885). Regardless –na surfaces as a result of hiatus repair in the language (cf. Albro 2005, Lin 2005, O’Neill 2015, a.o.).
³ ki-/ku- could be allomorphs of the same morpheme—but even if this is were true, their distributions do not appear to be conditioned by neighboring vowels (ki-: http://malagasyword.org/bins/derivLists?form=ki~#longScroll; ku-: http://malagasyword.org/bins/derivLists?form=ko~#longScroll)
References


