*Phonology* (2023), **40**:1-2 1–33
doi:10.1017/S0952675724000083

PHONOLOGY

ARTICLE

# A learning-based account of local phonological processes

Caleb A. Belth

Department of Linguistics, University of Utah, Salt Lake City, UT, USA.
Email: caleb.belth@utah.edu

**Received:** 5 November 2021; **Revised:** 24 August 2022; **Accepted:** 1 March 2023;
**First published online:** 22 October 2024

**Keywords:** learning; phonological processes; locality; computational modelling; alternations; abduction

**Abstract**

Phonological processes tend to involve local dependencies, an observation that has been expressed explicitly or implicitly in many phonological theories, such as the use of minimal symbols in SPE and the inclusion of primarily strictly local constraints in Optimality Theory. I propose a learning-based account of local phonological processes, providing an explicit computational model. The model is grounded in experimental results that suggest children are initially insensitive to long-distance dependencies and that as their ability to track non-adjacent dependencies grows, learners still prefer local generalisations to non-local ones. The model encodes these results by constructing phonological processes starting around an alternating segment and expanding outward to incorporate more phonological context only when surface forms cannot be predicted with sufficient accuracy. The model successfully constructs local phonological generalisations and exhibits the same preference for local patterns that humans do, suggesting that locality can emerge as a computational consequence of a simple learning procedure.

## Contents

**1. Introduction** **2**
1.1. Locality . . . 2
1.2. The nature of the learning task . . . 3
1.3. Locality and identity as principles of computational efficiency . . . 5
**2. Model: PLP** **5**
2.1. The input . . . 6
2.2. Constructing generalisations . . . 6
2.3. Encoding generalisations in a grammar . . . 10
2.4. Updating incrementally . . . 14
**3. Prior models** **15**
3.1. Constraint-based models . . . 15
3.2. Rule-based, neural network, and linear discriminative models . . . 15
3.3. Formal-language-theoretic models . . . 16
**4. Evaluating the model** **17**
4.1. Model comparisons . . . 17
4.2. Comparison to humans' preference for locality . . . 18
4.3. Learning German devoicing . . . 21
4.4. Learning a multi-process grammar . . . 24
4.5. Learning Tswana's post-nasal devoicing . . . 26
**5. Discussion** **27**
5.1. The nature of locality . . . 27
5.2. Future directions . . . 27

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

**A. PLP and strict locality** **28**
A.1. Strict locality of sequences . . . . . 28
A.2. Strict locality of generalisations . . . . . 28
**B. Differences between PLP and MGL** **29**
B.1. Generalisation strategy . . . . . 29
B.2. Number of rules . . . . . 29
B.3. Production . . . . . 29
B.4. Natural classes . . . . . 29

## 1. Introduction

Phonological processes tend overwhelmingly to involve dependencies between adjacent segments (Gafos 1999; Chandlee *et al.* 2014). For example, the English plural allomorph depends on the stem-final segment, to which it is adjacent, as in (1).

(1) /dɑg-z/ → [dɑgz]
/kæt-z/ → [kæts]
/hɔrs-z/ → [hɔrsəz]

Moreover, underlying forms are often posited to be minimally different from surface forms, exhibiting abstractness only when surface alternation necessitates it (Kiparsky [1968] 1982; Peperkamp *et al.* 2006; Ringe & Eska 2013; Richter 2021). This is supported by experimental findings, where children avoid introducing discrepancies between surface and underlying forms when there is little motivation for doing so (Jusczyk *et al.* 2002; Kerkhoff 2007; Coetzee 2009; van de Vijver & Baer-Henney 2014).

When, and only when, concrete representations are abandoned in favour of (minimally) abstract underlying representations, a child must learn a phonological process to derive the surface form from the abstract underlying form. Experimental studies are revealing about the mechanism underlying sequence learning: humans show a strong proclivity for tracking adjacent dependencies, beginning to track non-adjacent dependencies only when the data overwhelmingly demands it (Saffran *et al.* 1996, 1997; Aslin *et al.* 1998; Santelmann & Jusczyk 1998; Gómez 2002; Newport & Aslin 2004; Gómez & Maye 2005). As Gómez & Maye (2005: 199) put it, 'It is as if learners are attracted by adjacent probabilities long past the point that such structure is useful'. Indeed, artificial-language experiments have repeatedly demonstrated that learners more easily learn local phonological processes than non-local ones (Baer-Henney & van de Vijver 2012) and, when multiple possible phonological generalisations are consistent with exposure data, learners systematically construct the most local generalisation (Finley 2011; White *et al.* 2018; McMullin & Hansson 2019).

In this article, I hypothesise a mechanistic account of how learners construct phonological generalisations, modelling the learner's attention as initially fixed locally and expanding farther only when local dependencies do not suffice. The proposed model incorporates the idea that the learning of a phonological process is triggered when, and only when, underlying abstraction introduces discrepancies between underlying and surface representations (Kiparsky [1968] 1982). I view the model's locally centred attention and default assumption of identity as computationally parsimonious, and thus call it the *Parsimonious Local Phonology* (PLP) learner. When presented with small amounts of child-directed speech, PLP successfully learns local phonological generalisations. PLP's search strategy – starting as locally as possible – leads it to accurately exhibit the same preference for local patterns that humans do. Next, I review experimental results on locality in §1.1, the view of learning that PLP adopts in §1.2 and how these reflect principles of efficient computation in §1.3.

### *1.1. Locality*

Early studies of statistical sequence learning found infants to be sensitive only to dependencies between adjacent elements in a sequence. Saffran *et al.* (1996, 1997) and Aslin *et al.* (1998) found infants

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

as young as 8 months to be sensitive to dependencies between adjacent elements, while Santelmann & Jusczyk (1998) found that even at 15 months, children did not track dependencies between non-adjacent elements. Studies with older participants revealed that the ability to track non-adjacent dependencies does eventually emerge: adults show a sensitivity to dependencies between non-adjacent phonological segments (Newport & Aslin 2004), and 18-month-old children can track dependencies between non-adjacent morphemes (Santelmann & Jusczyk 1998). However, even as sensitivity to non-adjacent dependencies develops, learners still more readily track local dependencies. Gómez (2002) found that 18-month-olds could track non-adjacent dependencies, but that they only did so when adjacent dependencies were unavailable. Gómez & Maye (2005) replicated these results with 17-month-olds, and attempted to map the developmental trajectory of this ability to track non-adjacent dependencies, finding that it grew gradually with age. At 12 months, infants did not track non-adjacent dependencies, but they began to by 15 months, and showed further advancement at 17 months. These experiments involved a range of elements: words, syllables, morphemes and phonological segments. Moreover, similar results have been observed in different domains, such as vision (Fiser & Aslin 2002). Together, these results suggest that learners might discover only local patterns at early stages in development and that even after sensitivity to less local patterns emerges, a preference for local patterns persists.

Further experiments targeted phonological learning in particular. Subjects in Finley's (2011) artificial-language experiments learned bounded (local) harmony patterns and did not extend them to non-local contexts when there was no evidence for doing so. However, when exposed to unbounded (non-local) harmony patterns, subjects readily extended them to local contexts. This asymmetry suggests that learners will not posit less local generalisations until the evidence requires it. McMullin & Hansson (2019) replicated these results with patterns involving liquids and with dissimilation. Baer-Henney & van de Vijver (2012) used an artificial-language experiment to test the role of locality (as well as substance and amount of exposure) in learning contextually determined allomorphs. They found that when the allomorph was determined by a segment two positions away, learners more easily acquired and extended the pattern than when the allomorph was determined by a segment three positions away. In short, these studies demonstrate that learners posit the most local generalisation consistent with the data.

### *1.2. The nature of the learning task*

I adopt the view of others (e.g., Hale & Reiss 2008; Ringe & Eska 2013; Richter 2021) that children initially store words concretely, as accurately to what they perceive as their representational capacities allow. As their lexicon grows, surface alternations sometimes motivate the positing of abstract underlying forms, which introduce discrepancies between underlying and surface forms. For example, as Richter (2018, 2021) has characterised in rigorous detail, alternations such as 'eat' [it] ~ 'eating' [iɾɪŋ] lead to the flap [ɾ] and stop [t] being collapsed into allophones of underlying /T/. Similarly, a morphemic surface alternation such as 'cats' [kæts] ~ 'dogs' [dɑgz] may motivate an abstract underlying plural suffix /-Z/ (or default /-z/; Berko 1958). This view is in the spirit of Kiparsky's ([1968] 1982) Alternation Condition, and has been termed *invariant transparency* (Ringe & Eska 2013).

A consequence is that when, and only when, concrete segments are collapsed into abstract underlying representations, the need for a phonological grammar arises, to derive the surface forms for abstract underlying forms. I will use the example of stops following nasals to exemplify two significant corollaries. Voiceless stops following nasals are often considered to be a marked sequence, because post-nasal articulation promotes voicing, and post-nasal voicing is typologically pervasive (Locke 1983; Rosenthall 1989; Pater 1999; Hayes & Stivers 2000; Beguš 2016, 2019). Nevertheless, many languages – for example, English – tolerate post-nasal voiceless stops,[1] and a few even exhibit productive,

[1] I note that passive, phonetic post-nasal voicing still occurs in some such languages (Hayes & Stivers 2000); I am referring here to phonological voicing.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

phonological post-nasal *devoicing*. For example, Coetzee & Pretorius (2010) performed a detailed experimental study of Tswana speakers, finding that some extended post-nasal devoicing, as in (2), productively to nonce words.

(2) *Post-nasal devoicing in Tswana* (Coetzee & Pretorius 2010: 406)

| | | | | |
|---|---|---|---|---|
| a. | /m-batla/ | → | [mpatla] | 'want me' |
| | /m-botsa/ | → | [mpotsa] | 'ask me' |
| | /m-bulela/ | → | [mpulela] | 'open (for) me' |
| b. | /re-batla/ | → | [rebatla] | 'want us' |
| | /re-botsa/ | → | [rebotsa] | 'ask us' |
| | /re-bulela/ | → | [rebulela] | 'open (for) us' |

Beguš (2019: 699) found post-nasal devoicing to be reported as a sound change in 13 languages and dialects, and argues that although this pattern appears to operate against phonetic motivation, it likely emerged in each case as the result of a sequence of sound changes that were individually phonetically motivated.

Including a constraint to mark post-nasal voiceless stops in languages that tolerate them makes the learning task unnecessarily difficult, because the constraint must then be downranked despite the absence of surface alternations. Instead, under invariant transparency, children learning languages that tolerate post-nasal voiceless stops will simply not learn a phonological process regarding post-nasal stops, because there is nothing to learn. Moreover, when surface alternations that lack or operate in opposition to phonetic motivation (e.g., post-nasal devoicing) occur synchronically due to diachronic processes or other causes, no serious problem arises: the child simply learns a phonological process to account for the observed alternation, as has been observed in experiments (Seidl & Buckley 2005; Beguš 2018).

The view that children initially hypothesise identity between surface and underlying forms enjoys experimental support. Jusczyk *et al.* (2002) found that 10-month-old infants better recognise faithful word constructions than unfaithful ones. Van de Vijver & Baer-Henney (2014) found that both 5–7-year-olds and adults were reluctant to extend German alternations to nonce words, preferring instead to treat the nonce SRs as identical to their URs. Kerkhoff (2007) reports a consistent preference for non-alternation in Dutch children aged 3–7 years. In an artificial-language experiment, Coetzee (2009) found that learners more often extend non-alternation than alternation to test words, suggesting that this is learners' default.

Of course, children's initial productions are not faithful to adult productions (Smith 1973; Fikkert 1994; Grijzenhout & Joppen 1998; Grijzenhout & Joppen-Hellwig 2002; Freitas 2003), but this is likely due to underdeveloped control of the child's articulatory system, rather than an early state of the adult grammar (see Hale & Reiss 2008, §3.1 for a detailed argument). For instance, children systematically fail to produce complex CC syllable onsets in early speech even in languages that allow complex onsets, like Dutch, German, Portuguese and English (Fikkert 1994; Grijzenhout & Joppen 1998; Grijzenhout & Joppen-Hellwig 2002; Freitas 2003; Gnanadesikan 2004). Clusters tend to be reduced by deleting a consonant, and development proceeds from a cluster reduction stage to a full CC production stage, suggesting the discrepancy may be due to limited articulatory control.

PLP is a model of how phonological processes are learned once underlying abstraction leads to discrepancies in (UR, SR) pairs, which constitute PLP's input. As some reviewers of this article pointed out, the task of learning phonological processes to account for discrepancies between underlying and surface forms is intertwined with the task of figuring out when such abstract underlying representations are formed, and what they are like. This is evident when comparing the English plural voicing alternation (e.g., *cats* [kæt**s**] ~ *dogs* [dɑg**z**]) to the Dutch plural voicing alternation (e.g., [bɛ**t**] 'bed' ~ [bɛ**d**ən] 'beds'; Kerkhoff 2007: 1). English speakers show clear productive, rule-like behaviour (Berko 1958), while Dutch speakers' generalisation is less clearly rule-like (Ernestus & Baayen 2003; Kerkhoff 2007). The Dutch alternation is obfuscated by its interaction with other voicing alternations such as

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

assimilation (Buckler & Fikkert 2016, §2). Consequently, it may be that the English alternation is systematic enough to drive the learner to systematic underlying abstraction, while the Dutch alternation is not.

Thus, a complete theory of phonological learning must include, in addition to the mechanism by which processes are learned, a precise mechanism characterising how and when abstract underlying forms are posited. For example, Richter (2018, 2021) has hypothesised a mechanism by which learners abandon the null hypothesis of concrete underlying forms in favour of abstraction, and applied it to the case of the English [t] ~ [ɾ] allophones. The results closely matched lexical studies of child utterances, including a U-shaped development curve. Thus, PLP is just one part of the story. However, I believe that this part of the story – learning phonological processes from (UR, SR) pairs – is nevertheless important, and in line with the vast majority of prior work on learning phonological grammars, which have likewise tended to presuppose abstract underlying forms for use in, for example, constraint ranking (Legendre *et al.* 1990; Boersma 1997; Tesar & Smolensky 1998; Boersma & Hayes 2001; Smolensky & Legendre 2006; Boersma & Pater 2008).[2]

### *1.3. Locality and identity as principles of computational efficiency*

Locality and identity have natural interpretations as principles of computational efficiency, or 'third factors' (Chomsky 2005; Yang *et al.* 2017). The more local the context around an underlying segment, the fewer segments the cognitive system need be sensitive to in determining its output (Rogers *et al.* 2013: 99). Moreover, it is computationally simpler to copy input segments to the output unaltered than to change them in the process.

I present my proposed model in §2, discuss prior models in §3, evaluate the model in §4, and conclude with a discussion in §5.

## 2. Model: PLP

The proposed model is called PLP, for *Parsimonius Local Phonology* learner. PLP learns from an input of (UR, SR) pairs, which may grow over time as the learner's vocabulary expands. It constructs the generalisations necessary to account for which segments surface unfaithfully in those pairs and in what phonological contexts that happens. These generalisations are placed in a grammar, for use in producing output SRs for input URs.

(3) *PLP learning algorithm*
**Input:** (UR, SR) pairs
  a. Initialise an empty grammar $G$ and empty vocabulary $\mathcal{V}$
  b. **While** there are more pairs $(u, s)$ to learn from **do**
    i. Update $\mathcal{V}$ with $(u, s)$
    ii. Use $G$ to predict surface representation $\hat{s}$ for underlying $u$ (§2.3.4)
    iii. **For** each discrepancy between $u$ and $s$ not accounted for in $\hat{s}$ **do** (§2.1)
      α. Construct a generalisation $g$ for the discrepancy (§2.2)
      β. Encode $g$ in $G$ (§2.3)
    iv. Update any generalisations that now overextend due to $\mathcal{V}$ growth (§2.4)

[2] One reviewer pointed out that the concept of underlying forms faces scepticism and that many phonologists have rejected the concept altogether. I acknowledge that the view of learning described here is not uncontroversial. Hyman (2018) provides a discussion of the merits of underlying representations.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

PLP assumes identity between URs and SRs by default: it adds generalisations to *G* only in step (3b-iii), when discrepancies arise. A locality preference emerges from the generalisation strategy it employs in steps (3b-iii-α) and (3b-iv): PLP starts with the narrowest context around an unfaithfully surfacing segment and proceeds further outward from the segment only when an adequate generalisation cannot be found. Consequently, I consider steps (3b-iii-α) and (3b-iv), together with the addition of generalisations to the grammar only when motivated by discrepancies, to be PLP's main contributions. The code is available on GitHub.[3]

### *2.1. The input*

The input to PLP is a set of (UR, SR) pairs, which may grow over time, simulating the learner's vocabulary growth. As discussed in §1.2, discrepancies between a UR and its corresponding SR arise when a learner abandons concrete underlying representations in favour of underlying abstraction. A discrepancy can be an input segment that does not surface (deletion), an output segment that has no input correspondent (epenthesis) or an input segment with a non-identical output correspondant (segment change). In this work, I treat the (UR, SR) pairs, with discrepancies present, as PLP's input. Future work will combine this with the important problem of when abstract underlying forms are posited (e.g., Richter 2018). I also assume that the correspondence between input and output segments is known. The same assumption is tacit in constraint ranking models, which use the correspondence for computing faithfulness constraint violations.

The URs and SRs are sequences of segments, which I treat as sets of distinctive features (Jakobson & Halle 1956; Chomsky & Halle 1968). Thus, structuring sound into a phonological segment inventory organised by distinctive features is treated as a separate learning process (e.g., Mayer 2020). I use feature assignments from Mortensen *et al.* (2016).

I will use the English plural allomorph as a running example. Suppose that at an early stage in acquisition, a child has memorised some of the plural forms of nouns in their vocabulary, as in (4).

(4) /dɑgz/, /kæts/, /hɔrsəz/, . . .

At this stage, an empty grammar, which regurgitates each memorised word, will suffice. Moreover, since no discrepancies yet exist, PLP will be content with this empty grammar: the **for** loop (step (3b-iii)) will not be entered. As the child begins to learn morphology, they may discover the morphological generalisation that plurals tend to be formed by suffixing /-z/. All of the child's plural URs will then, in effect, be reorganised as in (5).

(5) /dɑg-z/, /kæt-z/, /hɔrs-z/, . . .

At this point, when the child goes to use their grammar (step (3b-ii)), they will discover that it now predicts *[kætz] and *[hɔrsz], inconsistent with their expectation based on prior experience with the words. The newly introduced discrepancies trigger the **for** loop (step (3b-iii)) and require PLP to provide an explanation for them. Suppose the first word to trigger this is /kæt-z/, erroneously predicted as *[kætz] instead of the expected [kæts]. PLP then constructs a generalisation to capture the phonological context in which /z/ surfaces as [s] (step (3b-iii-α)).

### *2.2. Constructing generalisations*

The core component of PLP is its component for constructing generalisations (step (3b-iii-α)).

[3] https://github.com/cbelth/PLP

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

*2.2.1. The structure of generalisations*

The generalisations that PLP constructs are pairs $g = (\bar{s}, a) \in \mathcal{S} \times \mathcal{A}$, where $\bar{s} \in \mathcal{S}$ (6) is a sequence and $a \in \mathcal{A}$ (7) is an action carried out at a particular position in the sequence. Each element in a sequence is a set of segments from the learner's segment inventory, $\Sigma$ (6).[4]

(6) $\mathcal{S} \triangleq \bigcup_{k=1}^{\infty} \{s_1 s_2 \ldots s_k : s_i \subset \Sigma\}$

A set of segments may be extensional, e.g., $s_i$ = {s, ʃ, z, ʒ}, or a natural class – e.g., $s_i$ = [+sib]. An action can be any of those listed in (7): deletion of the $i$th segment, insertion of new segment(s) to the right of the $i$th segment,[5] or setting the $i$th segment's feature $f$ to + or −.[6]

(7) $\mathcal{A} \triangleq \{\textsc{Del}(i), \textsc{Ins}(s_{\text{new}}, i), \textsc{Set}(f, \pm, i)\}$

For example, the generalisation in (8a) states that a consonant is deleted when it follows and precedes other consonants; (8b) says that a schwa is inserted to the right of any sibilant that precedes another sibilant; and (8c) says that the voicing feature of voiced obstruents in syllable-final position is set to − (using $]_\sigma \in \Sigma$ to indicate the right boundary of a syllable).

(8) a. ([+cons][+cons][+cons], Del(2))
b. ([+sib][+sib], Ins('ə', 1))
c. ([+voi, −son] $]_\sigma$, Set(voi, '−', 1))

Any grammatical formalism capable of encoding these generalisations could be used, but in this article I chose a rule-based grammar. The specified set of possible actions is meant to cover a majority of phonological processes, but more could be added if necessary (e.g. metathesis).

The part of the sequence picked out by the index $i$ determines the target of the rule, and the part of the sequence to the left and right of $i$ determine the rule's left and right contexts. Each type of action (7) can be encoded in one of the rule schemas in (9), where $k = |\bar{s}|$.

(9) a. $\textsc{Del}(i)$ $\quad s_i \rightarrow \varnothing \;/\; s_1 \ldots s_{i-1} \_\_ s_{i+1} \ldots s_k$
b. $\textsc{Ins}(s_{\text{new}}, i)$ $\quad \varnothing \rightarrow s_{\text{new}} \;/\; s_1 \ldots s_i \_\_ s_{i+1} \ldots s_k$
c. $\textsc{Set}(f, \text{'+'}, i)$ $\quad s_i \rightarrow [+f] \;/\; s_1 \ldots s_{i-1} \_\_ s_{i+1} \ldots s_k$
d. $\textsc{Set}(f, \text{'−'}, i)$ $\quad s_i \rightarrow [-f] \;/\; s_1 \ldots s_{i-1} \_\_ s_{i+1} \ldots s_k$

Thus, the generalisations in (8) are encoded as the rules in (10).

(10) a. [+cons] → ∅ / [+cons] __ [+cons]
b. ∅ → ə / [+sib] __ [+sib]
c. [+voi, −son] → [−voi] / __ $]_\sigma$

Each sequence in $\mathcal{S}$ is *strictly local* (McNaughton & Papert 1971) – describing a contiguous sequence of segments (see Appendix A for elaboration) – and has the same structure as the 'sequence of feature matrices' constraints from Hayes & Wilson (2008: 391). Moreover, the input–output relations described by each generalisation are probably[7] *input strictly local* maps (Chandlee 2014). These structures are not necessarily capable of capturing all phonological generalisations, and intentionally so. Typological considerations point to strict locality as a central property of generalisations, due to its prevalence (Chandlee 2014) and repeated occurrence across representations (Heinz *et al.* 2011). This article is intentionally targeting precisely those generalisations, and I discuss principled extensions for non-local generalisations in §5.2.

[4] These elements may also contain syllable- and word-boundary information, which I implement following Chomsky & Halle (1968) and Hayes & Wilson (2008) by introducing a [±segment] feature and corresponding [−segment] element in Σ to mark boundaries.

[5] Insertion in initial position is achieved with $i = 0$.

[6] More generally, we can treat the first parameter as a vector of features and the second as a vector of ± values to capture multiple feature changes, but for simplicity I only describe the case of a single feature change.

[7] It is generally believed that processes describable with the types of rules that PLP constructs are input strictly local maps (Chandlee 2014), but – to the best of my knowledge – there does not exist a published proof of this fact. See Appendix A for more.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

*2.2.2. Searching generalisations*

When PLP encounters a discrepancy – an input segment surfacing unfaithfully – it uses the algorithm in (11) to construct a generalisation $g = (\bar{s}, a)$. I refer to the discrepancy as $x \rightarrow y$, where $x$ is the input segment and $y \neq x$ is its surface realisation.

(11) *Algorithm for constructing generalisations*
**Input:** A discrepancy, $x \rightarrow y$, and the current training vocabulary $\mathcal{V}$
a. Initialise a window $\bar{w} = [\{x\}]$ of width one
b. Infer $a$ from $x \rightarrow y$ and initialise a generalisation $g = (\bar{w}, a)$
c. **While** $g$ is insufficiently accurate over $\mathcal{V}$ **do**[8] (§2.2.3)
   i. Expand the width of the window by length one (§2.2.4)
   ii. Set $g$'s sequence $\bar{s}$ to the most accurate context around $x$ that fits in $\bar{w}$ (§2.2.4)

PLP uses a window, $\bar{w}$, to control the breadth of its search. The window is a sequence of cells that can be filled in to create $g$'s sequence (6). The window starts with only one cell, filled with $s_0 = \{x\}$ (step (11a)). PLP then infers the type of change from $x$ to $y$ as in (12).

(12) $$a = \begin{cases} \textsc{Del} & \text{if } x \rightarrow \varnothing \;\; (y = \varnothing) \\ \textsc{Ins} & \text{if } \varnothing \rightarrow y \;\; (x = \varnothing) \\ \textsc{Set} & \text{otherwise} \end{cases}$$

For Ins, the value inserted ($s_{\text{new}}$) is $y$; for Set, the featured changed ($f$) and its value (+ or −) are inferred from the difference between $x$ and $y$. The index $i$ specifies where $x$ falls in $g$'s sequence, $\bar{s}$; initially, since $\bar{s} = \bar{w} = [\{x\}]$, $i = 1$ (step (11b)).

As Figure 1a visualises, PLP starts with the most local generalisation, which makes no reference to the segment's context: the segment always surfaces unfaithfully. In the running example, PLP first posits (13), which predicts that /z/ always surfaces as [s] (Figure 1b).

(13) /z/ → [−voi] / __

This, however, is contradicted by other words in the vocabulary: /z/ surfaces faithfully as [z] in words like [dɑgz] and with an epenthetic vowel in words like [hɔrsəz], which suggests that this initial generalisation is wrong (step (11c)) and that the breadth of the search must be expanded (Figure 1a).

*2.2.3. When to expand breadth of search*

To come to such a verdict, PLP computes the number of predictions the rule makes over the current vocabulary and how many of those are correct. The number of predictions (13) makes is the number of times /z/ appears in the learner's vocabulary, and those that surface as [s] are the correct predictions. There are a number of options for determining the adequacy of the generalisation. We could require a perfect prediction record, but this may be too rigid due to the near inevitability of exceptions in naturalistic data. More generally, we could place a threshold on the number or fraction of errors that the generalisation can make. The choice of criterion does not substantially change PLP: going from local generalisations to less local ones proceeds in the same way regardless of the quality criterion, which simply determines the rate at which the more local generalisations are abandoned. In this work, PLP uses the Tolerance Principle (Yang 2016) as the threshold, which states that a generalisation making $N$ predictions about what an underlying segment surfaces as is productive – and hence the **while** loop (step (11c)) can be exited – if and only if the number of incorrect predictions it makes (called $e$ for exceptions) satisfies (14).

(14) $e \leq \frac{N}{\ln N}$

[8]The loop also exits if the search runs out of context, in which case no sufficiently accurate generalisation is possible.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

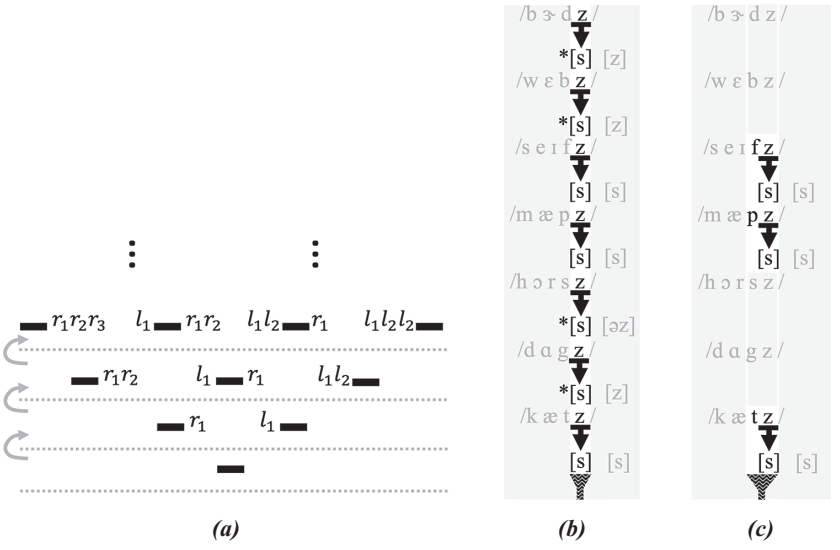



***Figure 1.*** *(a) The width of PLP's search expands outward (upward arrows) when and only when an adequate generalisation cannot be formed from a narrower context. (b) and (c) An example of PLP on seven English plural nouns. (b): PLP's first generalisation (13) is based on only the alternating segment and makes too many wrong predictions; this triggers PLP to expand its attention window. (c) PLP then forms generalisation (15a), which is based on the left-adjacent segment and allows the /z/ → [s] instances to be isolated.*

The threshold is cognitively motivated, predicting that children accept a linguistic generalisation when it is cognitively more efficient to do so (see Yang 2016, ch. 3 for how this threshold is identified). Since the threshold is based on cognitive considerations and has had success in prior work (e.g., Schuler *et al.* 2016; Koulaguina & Shi 2019; Emond & Shi 2021; Richter 2021; Belth *et al.* 2021), it is a reasonable choice for this article. In the current example, (13) has $N = 7$ and $e = 4$, which fails the criterion in (14): $4 > \frac{7}{\ln 7} (\approx 3.6)$. Thus, the **while** loop in (11c) is entered.

*2.2.4. Expanding breadth of search*

Once the initial hypothesis that /z/ always surfaces as [s] is ruled out as too errant, PLP adds one cell to the window (step (11c-i)). PLP fills the window with the sequence that matches the fewest of the sequences where /z/ does not surface as [s]. In other words, it chooses the context that better separates words like /kæt/ from words like /dɑg/ and /hɔrs/. Thus, for the vocabulary in Figure 1b and 1c, PLP prefers (15a) over (15b) because a left context of {t, p, f} is more successful than a right context of {#} at distinguishing the places where /z/ does indeed surface as [s] from those where it does not.[9] That is, PLP chooses the rule with the most accurate context fitting in the current window, where accuracy is measured as the fraction of the rule's predictions over the training URs that match the corresponding training SRs. In our example, then, PLP's second hypothesis is that /z/ surfaces as [s] whenever it follows a /t/, /p/, or /f/.

(15) a. /z/ → [−voi] / {t, p, f} __
b. /z/ → [−voi] / __ {#}

Figure 1a visualises PLP's search: it hypothesises a context where an underlying segment surfaces as some particular segment other than itself, checking whether the hypothesis is satisfactorily accurate,

[9] The symbol # denotes a word boundary.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

and expanding the breadth of its search if not. This process halts once a sufficiently accurate hypothesis has been discovered.

### *2.3. Encoding generalisations in a grammar*

The generalisations that PLP constructs are encoded in a grammar to be used in producing an SR for an input UR. The grammar, $G$, consists of a list of rules. Each time PLP constructs a generalisation (step (3b-iii-α)), it is placed in the appropriate rule schema (9) and added to the list of rules. If PLP replaces a generalisation due to underextension or overextension (step (3b-iv)), as described in §2.4, the old, offending rule is removed and a new one added. §2.3.1 discusses how rules that carry out the same action are combined; §2.3.2 discusses how natural classes are induced; §2.3.3 discusses how the list of rules is ordered and §2.3.4 discusses how $G$ produces outputs from inputs.

#### *2.3.1. Combining generalisations*

Generalisations that carry out the same change over different segments are combined in the grammar, so long as the resulting rule is accurate to a degree that satisfies the criterion in (14). For instance, the three rules in (16a) would be grouped into the single rule (16b).

(16) a. /d/ → [−voi] / __ ]$_\sigma$
/v/ → [−voi] / __ ]$_\sigma$
/g/ → [−voi] / __ ]$_\sigma$
b. {d, v, g} → [−voi] / __ ]$_\sigma$

#### *2.3.2. Inducing natural classes*

Up to this point, PLP's generalisations have been over sets of particular segments. Humans appear to generalise from individual segments to natural classes, as has been recognised by theory (Chomsky & Halle 1965; Halle 1978; Albright 2009) and evidenced by experiment (Berent & Lennertz 2007; Finley & Badecker 2009; Berent 2013).

PLP thus attempts to generalise to natural classes for each set of segments in a generalisation's sequence $\bar{s}$, in terms of shared distinctive features (Jakobson & Halle 1956; Chomsky & Halle 1968). The procedure can be thought of as retaining only the features shared by segments in $\bar{s}$ needed to keep the rule satisfactorily accurate. To exemplify this part of the model, I will assume PLP has constructed the epenthesis rule in (17), which produces mappings such as /hɔrsz/ → [hɔrsəz].

(17) ∅ → ə / {s, ʃ, z} __ {z}

The procedure, outlined in (18), starts with a new sequence $\bar{n}$ of length $|\bar{s}|$, with each element an empty natural class (step (18a)).

(18) *Procedure for inducing natural classes*
**Input:** A generalisation $g = (\bar{s}, a)$
a. Initialise a new generalisation $g_{\text{nc}} = (\bar{n}, a)$ with empty natural classes, $\bar{n}$
b. Initialise feature options for natural classes
c. **While** $g_{\text{nc}}$ is insufficiently accurate over $\mathcal{V}$ **do**
Add to $\bar{n}$ the feature that best narrows $\bar{n}$'s extension down to $\bar{s}$'s
d. Replace $g$ with $g_{\text{nc}}$

For the rule in (17), the sequence $\bar{s}$ is (19a) and the (empty) initial natural class sequence is (19b). Each element of $\bar{n}$ can take any feature shared by the corresponding segments in $\bar{s}$, so the set of feature options is (19c), which includes elements like (+sib, 1) because {s, ʃ, z} share '+sib' as a feature and (+voi, 2) because {z} has '+voi' as a feature, but it does not include (+voi, 1) because {s, ʃ, z} do not agree in this feature.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

(19) a. $\bar{s}$ = {s, ʃ, z}{z}
b. $\bar{n}$ = [ ] [ ]
c. {(+cons, 1), (+sib, 1), (−son, 1), ..., } ∪ {(+sib, 2), (+voi, 2) ...}

In the **while** loop (step (18c)), features are added one at a time to $\bar{n}$, choosing at each step the feature from (19c) that best narrows the extension of $\bar{n}$ (initially all sequences of length $|\bar{s}|$) to those in the extension of $\bar{s}$ (which is {sz, ʃz, zz}). Thus, adding the feature '+sib' to the first natural class (20a) will narrow $\bar{n}$'s extension towards $\bar{s}$'s better than '+cons'. As before, the new generalisation $g_{nc}$ is evaluated according to the Tolerance Principle threshold in (14). In the current example, $\bar{n}$ (20a) will still have sequences like {st, zi, ʃu, ...} in its extension, so '+sib' will then be added to the second natural class (20b).

(20) a. $\bar{n}$ = [+sib] [ ]
b. $\bar{n}$ = [+sib] [+sib]

This new sequence, $\bar{n}$, still has an extension greater than the original $\bar{s}$. However, because adjacent sibilants are indeed disallowed in English, this inductive leap is possible, and thus (17) will be replaced with (21) in the grammar.

(21) ∅ → ə / [+sib] __ [+sib]

This differs from the natural class induction in Albright & Hayes (2002, 2003), which generalises as conservatively as possible by retaining all shared features (see §B.4).

It may be possible for natural class induction to influence rule-ordering, so PLP identifies natural classes before determining the order in which the rules should apply. Specifically, natural classes are induced with rules temporarily ordered by scope (narrowest first), before the final ordering is computed as in §2.3.3.

### *2.3.3. Rule ordering*

In some cases, phonological processes may interact, in which case the interacting rules may need to be ordered. The topic of rule interaction and ordering has received immense attention in the literature – especially in discussions of opacity – and extends well beyond the scope of the current article. However, I will summarise PLP's approach to rule ordering, and characterise the path to a more systematic study of PLP's handling of complex rule interactions.

The standard rule interactions discussed in the literature are FEEDING, BLEEDING, COUNTERFEEDING and COUNTERBLEEDING, described in (22) following McCarthy (2007) and Baković (2011).

(22) Given two rules $r_i$ and $r_j$, where $r_i$ precedes $r_j$,
a. $r_i$ **feeds** $r_j$ iff $r_i$ creates additional inputs to $r_j$
b. $r_i$ **bleeds** $r_j$ iff $r_i$ destroys potential inputs to $r_j$
c. $r_j$ **counterfeeds** $r_i$ iff $r_j$ creates additional inputs to $r_i$
d. $r_j$ **counterbleeds** $r_i$ iff $r_j$ destroys additional inputs to $r_i$

Counterfeeding and counterbleeding are counterfactual inverses of feeding and bleeding: if $r_j$ counterfeeds (or counterbleeds) $r_i$, it would feed (or bleed) $r_i$ if it preceded $r_i$. McCarthy's (2007, §5.3) example of feeding, reproduced in (23), comes from Classical Arabic, where vowel epenthesis before word-initial consonant clusters ($r_i$) feeds [ʔ] epenthesis before syllable-initial vowels ($r_j$).

(23) *Feeding order in Classical Arabic* (McCarthy 2007: 103)

| | | |
|---|---|---|
| Underlying | /dˤrib/ | 'beat!' (MASC.SG) |
| Vowel epenthesis | idˤrib | |
| ʔ-epenthesis | ʔidˤrib | |
| Surface | [ʔidˤrib] | |

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

McCarthy (2007, §5.4) also provides an example of counterfeeding. In Bedouin Arabic, short high vowels are deleted in non-final open syllables, and /a/ is raised in the same environment. However, as (24) shows, because deletion precedes raising, the raising of the short vowel /a/ to [i] does not feed deletion.

(24) *Counterfeeding order in Bedouin Arabic* (McCarthy 2007: 107)

| | | |
|---|---|---|
| Underlying | /dafaʕ/ | 'he pushed' |
| Deletion | — | |
| Raising | difaʕ | |
| Surface | [difaʕ] | |

Examples of bleeding and counterbleeding come from dialects of English where /t/ and /d/ are flapped – [ɾ] – between stressed and unstressed vowels, while /aɪ/ and /aʊ/ raise to [ʌɪ] and [ʌʊ] before voiceless segments. The canonical case is counterbleeding order, where raising occurs before underlying /t/ even when it surfaces as voiced [ɾ] on the surface, as in (25).

(25) *Counterbleeding order in English*

| | | |
|---|---|---|
| Underlying | /raɪtɚ/ | *writer* |
| Raising | rʌɪtɚ | |
| Flapping | rʌɪɾɚ | |
| Surface | [rʌɪɾɚ] | |

In less-discussed dialects of English in Ontario (Joos 1942) and in Fort Wayne, Indiana (Berkson *et al.* 2017), the flapping of voiceless /t/ as voiced [ɾ] bleeds raising as in (26).

(26) *Bleeding order in English*

| | | |
|---|---|---|
| Underlying | /raɪtɚ/ | *writer* |
| Flapping | raɪɾɚ | |
| Raising | — | |
| Surface | [raɪɾɚ] | |

Given two interacting rules $r_i$ and $r_j$, it is straightforward to order them by following standard arguments. Specifically, ordering $r_i$ before $r_j$ (feeding/bleeding order) will produce errors on data from a language where $r_j$ in fact precedes $r_i$ (counterfeeding/counterbleeding) and vice versa. For example, if we call English dialects where flapping counterbleeds raising (25) 'Dialect A' and the dialects with bleeding (26) 'Dialect B' (following Joos 1942), ordering flapping before raising in Dialect A will erroneously cause /raɪtɚ/ to surface as [raɪɾɚ] instead of [rʌɪɾɚ]. Consequently, the correct counterfeeding order will yield higher accuracy than feeding order for a leaner exposed to Dialect A. A symmetrical argument holds for ordering in Dialect B.

Thus, for each pair of learned rules, PLP chooses the pairwise ordering with higher accuracy. To yield a full ordering of the rules, PLP constructs a directed graph where each rule in $\mathcal{R}$ forms a node. PLP considers each pair of rules $(r_i, r_j) \in \mathcal{R} \times \mathcal{R}$ and places a directed edge from $r_i$ to $r_j$ iff the accuracy of $r_j \circ r_i$ (i.e., applying $r_i$ first and $r_j$ to its output) is greater than that of the reverse, $r_i \circ r_j$. The directed graph is then topologically sorted to yield a full ordering.[10] Rules that interact are assigned the order that achieves higher accuracy, and non-interacting rules are ordered arbitrarily.

The bigger challenge is the possibility that the interactions between $r_i$ and $r_j$ obfuscate the independent existence of the rules, thereby making it difficult for them to be discovered in the first place. Counterfeeding and counterbleeding present no issues, because applying each rule independently, directly over the UR, produces the same SR as applying them sequentially in counterfeeding/counterbleeding order. For example, in McCarthy's (2007) Bedouin Arabic example in (24), /a/ → [i] is accounted for by the raising rule, and there is no deletion in /dafaʔ/ → [difaʔ] to hinder

[10] A *topological sort* of a directed graph is a linear ordering of its nodes such that every ordering requirement encoded in its edges is preserved (Cormen *et al.* 2009, 612).

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

the discovery of the deletion rule. Similarly, the /a/ → [ʌ] discrepancy in (25) can be accounted for by raising without reference to flapping, and the /t/ → [ɾ] discrepancy can be accounted for by flapping without reference to raising. I give an empirical demonstration of PLP learning rules in counterbleeding order in §4.3.4.

Since bleeding destroys contexts where a rule would have applied, it can cause overextensions. For example, when PLP is attempting to construct a raising rule for (26), the rule in (27) (treating the diphthong as a single segment) would overextend to /raɪtɚ/.

(27) aɪ → ʌɪ / __ [−voi]

However, since PLP allows some exceptions in accordance with the Tolerance Principle, this will only matter if the bled cases are pervasive enough to push the rule over the threshold in (14). Whether this happens must be determined on a case-by-case basis by the learner's lexicon. If the threshold of exceptions is crossed, PLP will simply expand the width of its search. When flapping bleeds raising (26), raising occurs distributionally before underlying voiceless segments that are not between a stressed and an unstressed vowel. The latter condition describes the contexts where raising is not bled, and still falls within a fixed-size window of the raising target, such as the underlined portion of /raɪtɚ/. The general point here is that if two rules interact extensively, there is still likely to be a fixed-length context – possibly a slightly larger one – that accounts for the processes. In fact, Chandlee *et al.* (2018) have shown that a wide range of phonological generalisations characterised as opaque in the literature can be formalised as input strictly local maps. In Appendix A, I show that the rules PLP learns correspond to Input Strictly Local maps. Thus, I am optimistic that PLP can succeed even with instances of opaque rule interactions. §4.4 provides an empirical demonstration of PLP learning rules in bleeding order.

Feeding may require small adaptions to PLP. In (23), no issue arises for the vowel-epenthesis rule, which does the feeding. The search for a rule to account for epenthetic [ʔ] will proceed analogously to the bleeding case. There are two underlying environments where epenthetic [ʔ] surfaces: before underlyingly initial vowels (# __ V) and before underlying initial consonant clusters (i.e., where raising feeds epenthesis, # __ CC). These are disjoint contexts, so it may be appropriate to adapt PLP to allow it to return two disjoint rules from its search in (11) to account for a discrepancy. In that case, the rules in (28) account for ʔ-epenthesis directly from URs.

(28) ∅ → ʔ / # __ CC ('fed' ʔ-epenthesis cases)
∅ → ʔ / # __ V

Alternatively, PLP could be adapted such that the search for new generalisations (step (3b-iii-α)) operates over intermediate representations – specifically those derived by existing rules – instead of underlying representations. In that case, the ʔ-epenthesis rule could be directly learned over the intermediate forms derived by the vowel-epenthesis rule.

In summary, this article is not an attempt to provide a complete account of rule ordering, which is beyond its scope. The results in §4.3.4 and §4.4 provide empirical demonstration of PLP learning some interacting rules, and the above discussion provides an outline of how PLP approaches rule interaction and what extensions may be necessary.

### *2.3.4. Production*

The rules are applied one after another in the order produced by the procedure in §2.3.3. Each individual rule is interpreted under *simultaneous application* (Chomsky & Halle 1968), which means that when matching the rule's target and context, only the input is accessible, not the result of previous applications of the rule. Thus, following the example from Chandlee *et al.* (2014: 37), the rule in (29) applied simultaneously to the input string *aaaa* yields the output *abba* rather than *abaa*, because the second application's context is not obscured by the first application.

(29) *a* → *b* / *a* __ *a*

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

Simultaneous application is the interpretation of rules that corresponds to input-strictly local maps, as discussed in §A.2. Other types of rule application, such as iterative or directional (e.g., Howard 1972; Kenstowicz & Kisseberth 1979), could be used in future work.

Thus, for an input $u$ and ordered list of rules $\mathcal{R} = r_1, r_2, ..., r_{|\mathcal{R}|}$, the grammar's output $\hat{s}$ is given by the composition of rules in (30).

(30) $\hat{s} = G(u) = r_{|\mathcal{R}|} \circ r_{|\mathcal{R}|-1} \circ ... \circ r_1(u) = r_{|\mathcal{R}|}(r_{|\mathcal{R}|-1}(...r_1(u)))$

### *2.4. Updating incrementally*

As PLP proceeds, vocabulary growth may cause the grammar to become stale and underextend or overextend, at which point PLP updates any problematic generalisations (step (3b-iv)).

Denoting the discrepancies between the input $u$ and the predicted output $\hat{s}$ as $d(u, \hat{s})$, and those between $u$ and $s$ as $d(u, s)$, underextensions are defined in (31a) as discrepancies between the input and expected output that are not accounted for in PLP's prediction $\hat{s}$, and overextensions are defined in (31b) as discrepancies in the predicted output that should not be there. Here the symbol \ denotes set difference, and $\triangleq$ means 'equals by definition'.

(31) a. $U \triangleq d(u, s) \setminus d(u, \hat{s})$
b. $O \triangleq d(u, \hat{s}) \setminus d(u, s)$

Underextensions are handled by the **for** loop (step (3b-iii)). Inside the loop, a new generalisation is created (step (3b-iii-α)). This is encoded in the grammar (step (3b-iii-β)) by adding it to this list of rules. If a prior generalisation for the discrepancy exists, it is deleted from the list. An example of this is (32), where the word /mæp-z/ (32b) freshly enters the vocabulary.

(32) a. /dɑg-z/ → [dɑgz]
/kæt-z/ → [kæts]
/hɔrs-z/ → [hɔrsəz]
b. /mæp-z/ → [mæps]

Prior to its arrival, the rule (33a) was sufficient to explain when /z/ surfaces as [s]. This, however, fails to account for the new word, which ends in /p/ not /t/. PLP handles this by discarding the old rule and replacing it with a fresh one, such as (33b), derived by the same process described above in §2.2.

(33) a. /z/ → [−voi] / {t} __
b. /z/ → [−voi] / {t, p} __

Overextension – a discrepancy between the input $u$ and PLP's prediction $\hat{s}$ that did not exist between $u$ and the expected output $s$ – is handled by (step (3b-iv)). An example is (34), where (34b) enters the learner's vocabulary after (34a).

(34) a. /kæt-z/ → [kæts]
b. /dɑg-z/ → [dɑgz]

In such a case, the rule in (35) will have been sufficient to explain (34a), but will result in an erroneous *[dɑgs] for (34b).

(35) /z/ → [−voi] / __

PLP resolves this by discarding the previous rule and replacing it with a new one by the process in §2.2.

For both underextension and overextension, when the list of rules is updated, the steps in §2.3 – combining generalisations, inducing natural classes and ordering rules – are repeated. Since PLP can replace generalisations as needed as the vocabulary grows, it can learn incrementally, in batches, or once and for all over a fixed vocabulary.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

## 3. Prior models

### *3.1. Constraint-based models*

Constraint-ranking models rank a provided set of constraints. Tesar & Smolensky's (1998) Constraint Demotion algorithm was an early constraint-ranking model for OT. Others are built on stochastic variants of OT or Harmonic Grammar (HG; Legendre *et al.* 1990; Smolensky & Legendre 2006), including the Gradual Learning Algorithm (Boersma 1997; Boersma & Hayes 2001) for Stochastic OT and a later model (Boersma & Pater 2008) that provided a different update rule for HG (see Jarosz 2019 for an overview).

Constraint-ranking models can capture the assumption of classical OT that learning amounts to ranking a universal constraint set, or they can rank a learned constraint set. Hayes & Wilson's (2008) Maximum Entropy model learns and ranks constraints, but it learns phonotactic constraints over surface forms, not alternations as PLP does.

Locality and identity biases are better reflected in the content of the constraint set than in the constraint ranking algorithm. Locality is determined in virtue of what segments are accessed in determining constraint violations.

Constraint-ranking models usually begin with markedness constraints outranking faithfulness constraints (Smolensky 1996; Tesar & Smolensky 1998; Jusczyk *et al.* 2002; Gnanadesikan 2004). Consequently, any UR will initially undergo any changes necessary to avoid marked structures, even in the absence of surface alternations that would motivate discrepancies. Ranking faithfulness constraints above markedness constraints has been advocated by Hale & Reiss (2008), but this approach has not been widely adopted. This in part due to arguments that such an initial ranking would render some grammars unlearnable (Smolensky 1996), and in part due to the view that features of early child productions, in particular 'emergence of the unmarked', reflect an early stage of the child's grammar, rather than underdeveloped articulatory control.

### *3.2. Rule-based, neural network, and linear discriminative models*

Johnson (1984) proposed an algorithm for learning ordered rules from words arranged in paradigms as a proof of concept about the learnability of ordered-rule systems. This algorithm does not incorporate a locality bias and has not been extensively studied empirically or theoretically.

Albright & Hayes (2002, 2003) developed a model for learning English past tense morphology through probabilistic rules. The model can be applied to learn rules for any set of input–output word pairs, including phonological rules. It is called the Minimum Generalisation Learner, because when it seeks to combine rules constructed for multiple input–output pairs, it forms the merged rule that most tightly fits the pairs. A consequence of this generalisation strategy is that the phonological context of the rule is as wide as possible around the target segment, only localising around the target when less local (and hence less general) contexts cannot be sustained. This is the direct opposite of PLP and of experimental results that suggest human learners start with local patterns and move to non-local patterns only when local generalisations cannot be sustained (Finley 2011; Baer-Henney & van de Vijver 2012; McMullin & Hansson 2019). I further discuss differences between PLP and Minimal Generalisation Learner (MGL) in Appendix B.

Rasin *et al.* (2018) propose a Minimum Description Length model for learning optional rules and opacity. The authors intended the model as a proof of concept and only evaluated it on two small artificial datasets.

Peperkamp *et al.* (2006) propose a statistical model for learning allophonic rules by finding segments with near-complementary distributions. The method is not applicable to learning rules involving non-complementary distributions. Calamaro & Jarosz (2015) extend the model to handle some cases of non-complementary distributions, if the alternation is conditioned by the following segment (i.e., $a \rightarrow b / \_\_\ c$ where $|a| = |b| = |c| = 1$). These works attempt to model the very early stage of learning

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

alternations (White *et al.* 2008) prior to most morphological learning, whereas PLP models learning after abstract URs have begun to be learned.

Beguš (2022) trained a generative, convolutional neural network on audio recordings of English-like nonce words, which followed local phonological processes and a non-local process (vowel harmony). The model was then used to generate speech. This model-generated speech followed the local processes more frequently than the non-local process, suggesting that it more easily learned local than non-local processes. This is possibly due to the use of convolution, which is a fundamentally local operation. As a model for generating artificial speech, it is not directly comparable in the context of learning processes that map URs to SRs.

In a different direction, Baayen *et al.* (2018, 2019) propose using Linear Discriminative Learning to map vector representations of form onto vector representations of meaning and vice versa. Since this model operates over vector representations of form and meaning, it is not directly comparable.

### *3.3. Formal-language-theoretic models*

Formal-language- and automata-theoretic approaches analyse phonological generalisations in computational terms. Many resulting learning models attempt to induce a finite-state transducer (FST) representation of the map between SRs and URs. These automata-theoretic models, together with precise assumptions about the data available for learning, allow for learnability results in the paradigm of identification in the limit (Gold 1967). Such results state that a learning algorithm will converge on a correct FST representation of any function from a particular family, provided that the data presented to it meet certain requirements – called a *characteristic sample*. In phonology, the target class of functions is usually one that falls in the sub-regular hierarchy (Rogers *et al.* 2013), which contains classes of functions more restrictive than the regular region of the Chomsky Hierarchy (Chomsky 1959). These models are often chosen to demonstrate theoretical learnability results, and have seldom been applied to naturalistic data.

Gildea & Jurafsky (1996) developed a model, based on OSTIA (Oncina *et al.* 1993), which learns subsequential FSTs. The class of subsequential functions is a sub-regular class of functions that may be expressive enough to capture any type of observed phonological map (Heinz 2018), although some tonal patterns appear to be strong counterexamples (Jardine 2016). The authors intended their model only as a proof of concept of the role of learning biases, and it requires unrealistic quantities of data to learn effectively. Indeed, the authors recognise the importance of faithfulness and locality as learning biases, which they attempted to embed into OSTIA. Their biases were, however, heuristics. In particular, a bias for locality was introduced by augmenting states with the features of their neighbouring contexts. This in effect restricts the learner to local patterns, which is different from the current article's proposal, in which locality is a consequence of how the algorithm proceeds over hypotheses.

As Chandlee *et al.* (2014) observes, a more principled means of incorporating a locality bias into a finite-state model is to directly target the class of strictly local functions. Chandlee *et al.* (2014) propose such a model, called ISLFLA, and prove that it can learn any strictly local function in the limit, in the sense of Gold (1967). However, the characteristic sample for the algorithm includes the set of input–output pairs for every language-theoretically possible string up to length $k$ (a model-required parameter). As Chandlee *et al.* (2014) discusses, this is problematic, since natural language may in principle never provide all logically possible strings, due to phonotactic or morophological constraints. I implemented ISLFLA and attempted to run it on naturalistic data, and it does indeed fail to identify any FST on such data.[11]

Jardine *et al.* (2014) propose a model, SOSFIA, for learning subsequential FSTs when the FST structure is known in advance; only the output for each arc in the FST needs to be learned. Strictly

[11] OSTIA will run on data not satisfying its characteristic sample; it is just not guaranteed to induce a correct FST in such cases. In contrast, ISLFLA is unable to proceed if the characteristic sample is not met: it exits at line 9 of the pseudocode in Chandlee *et al.* (2014: 499).

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

local functions are such a case, because the necessary and sufficient automata-theoretic conditions of strict locality include a complete FST structure (Chandlee 2014). SOSFIA also admits learnability in the limit results but has not been applied to naturalistic data.

## 4. Evaluating the model

This section evaluates PLP, addressing the questions listed in (36):

(36) **Q1**. Does PLP reflect human learners' preference for local generalisations?
**Q2**. How well does PLP learn local generalisations?
**Q3**. What are the learning effects of assuming UR–SR identity by default?

### *4.1. Model comparisons*

I compare PLP to several alternative models.

#### *4.1.1. Rule-based, neural network and finite-state models*

**MGL** is the Minimal Generalisation Learner from Albright & Hayes (2002, 2003). I used the Java implementation provided by the authors. MGL may produce multiple candidate SRs for a UR if more than one rule applies to the UR. In such cases, I used the rule with the maximum conditional probability scaled by scope (*confidence* in the terminology of Albright & Hayes 2002, §3.2) to derive the predicted SR.

**ED** (encoder–decoder) is a neural network model. It is a successful neural network model for many natural language processing problems involving string-to-string functions, such as machine translation between languages (Sutskever *et al.* 2014) and morphological reinflection (Cotterell *et al.* 2016). It has also been used to revisit the use of neural networks in the 'past tense debate' of English morphology (Kirov & Cotterell 2018), though its use as a computational model of morphological acquisition has been called into question (McCurdy *et al.* 2020; Belth *et al.* 2021). I follow Kirov & Cotterell (2018) and Belth *et al.* (2021) in its setup, using the same RNN implementation, trained for 100 epochs, with a batch size of 20, optimising the log-likelihood of the training data. Both the encoder and the decoder are bidirectional long short-term memory networks (LSTMs) with 2 layers, 100 hidden units, and a vector size of 300.

**OSTIA** (Oncina *et al.* 1993) is a finite-state model for learning subsequential finite-state transducers. I used the Python implementation from Aksënova (2020).

**ID** is a trivial baseline that simply copies every input segment to the output. This allows for interpreting the value of assuming UR–SR identity by default.

#### *4.1.2. Learning as constraint ranking*

I also compare PLP to the view of learning as ranking a provided constraint set. Classic OT views constraints as part of UG; I represent this view with **UCON**, for *universal constraint set*. An alternative view is that the constraint set is learned; I represent this view with **ORACLE**, which effectively constitutes an upper bound on how well a model that learns the constraint set to be ranked could do. ORACLE is provided with all and only the markedness constraints relevant to the grammar being learned. UCON is provided with the same constraints as ORACLE, plus two extra markedness constraints that are violable in the adult languages and thus must be downranked.

It is important to emphasise that these models learn in a different setting from PLP and those in §4.1.1. The latter receive as input only UR–SR training pairs, whereas UCON and ORACLE receive both training pairs and a constraint set. Consequently, UCON and ORACLE's accuracy levels in producing SRs are not directly comparable to those of other models. My goal in comparing PLP to UCON and ORACLE is to highlight how PLP's account of phonological learning differs from theirs.

For UCON and ORACLE, I use the Gradual Learning Algorithm (GLA; Boersma 1997; Boersma & Hayes 2001) to rank the constraints, because it is robust to exceptions – an important property when

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

learning from noisy, naturalistic data. I emphasise, however, that the comparison is not to the particular constraint-ranking algorithm; others could have been chosen. Because the experiments involve many random samples and tens of thousands of tokens, the implementation of GLA in Praat (Boersma 1999) was not well-suited. Thus, I used my own Python implementation of GLA, with the same default parameters as in Praat (evaluation noise: 2.0, plasticity: 1.0). I initialise markedness constraints above faithfulness constraints.

### *4.2. Comparison to humans' preference for locality*

In an experimental study, Baer-Henney & van de Vijver (2012) found that allomorphic generalisations in an artificial language were more easily and successfully learned when the surface allomorph was determined by a segment two positions away than when it was determined by a segment three positions away. The study involved three artificial languages in which plural nouns were formed by affixing either the [−back] vowel [-y] or its [+back] counterpart [-u]. Each language involved a different phonological condition for determining which affix surfaced. Treating /-y/ as the underlying affix, the three generalisations are those in (37).

(37) a. [−back] → [+back] / [+vowel,+**back**][+cons] __
b. [−back] → [+back] / [+vowel,+**tense**][+cons] __
c. [−back] → [+back] / [+cons,+**son**][+vowel][+cons] __

All singular forms are CVC words; plurals add a vowel. The (37a) language is an example of vowel harmony, since the affix vowel assimilates in backness to the preceding vowel. The (37b) language is equally local, but lacks clear phonetic motivation, since the stem vowel feature that determines the affix's backness is [tense]. The (37c) language is both less local and phonetically unmotivated, since the backness of the suffix vowel is determined by the initial consonant of the stem. Because all three languages have CVC stems and CVCV plurals, each pattern is strictly local, but (37a) and (37b) each involve a sequence of three contiguous segments, while (37c) involves four.

Since PLP starts locally around the affix when looking for an appropriate generalisation, and only proceeds outward when the more local contexts become too inaccurate, I expect PLP to learn the (37a) and (37b) generalisations substantially more easily than the (37c) generalisation, just as Baer-Henney & van de Vijver (2012) found for humans (**Q3**). For comparison, I use MGL, which generalises in roughly the opposite way: it constructs the narrowest – and hence less local – generalisation. I also compare to grammars resulting from ranking three different constraint sets. The markedness constraints for (37) are listed in (38).

(38) a. *[+vowel, +back][+cons][−back, +vowel]
b. *[+vowel, +tense][+cons][−back, +vowel]
c. *[+cons, +son][+vowel][+cons][−back, +vowel]

The first constraint set encodes the assumption of a universal constraint set containing only grounded, universal constraints by including only (38a), because it is the only generalisation viewed as phonetically motivated. Second, I consider a constraint set containing all three markedness constraints in (38) regardless of which language is being learned. Third, I consider a constraint set containing only the constraint relevant to the language being learned.

Baer-Henney & van de Vijver (2012) found not only that the local generalisations were learned more easily than the non-local one, but also that the phonetically motivated generalisation (37a) was learned slightly more easily than (37b). The authors argued that this is evidence for substantive bias in phonological learning. However, the question of substantive bias is largely orthogonal to the current article, since my focus is on locality. Moreover, the performance gap between (37a) and (37b) was much smaller than the gap between them and (37c). For these reasons, I focus on the difference in models' performance on (37a,b) *vs.* (37c) in this experiment.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

*4.2.1. Setup*

Each of Baer-Henney & van de Vijver's (2012) experiments involved presenting subjects with randomly selected singulars and plurals from the respective artificial languages. Each word was accompanied by a picture conveying the word's meaning; one item was present in the picture for singulars and multiple items for plurals. The singulars and plurals were presented independently, so the experimental setup did not separate phonological learning from learning the artificial languages' morphology and semantics. Because of this, the study participants likely only successfully acquired the underlying and surface representations for a subset of the exposure words, and what fraction of the exposure set they learned is entirely unknown. Since the models assume URs and SRs as training data, I factor out the fraction of the exposure set for which they have acquired URs and SRs by treating it as a free variable that the models can optimise over. I use the data released by Baer-Henney & van de Vijver (2012) and follow their setup to construct training (exposure) and test sets.[12] I ran each model over 100 randomised exposure sets to simulate 100 participants.

The MGL model from Albright & Hayes (2002, 2003) combines rules that target the same segment and carry out the same change to that target. For instance, if it has acquired the two word-specific rules in (39a) and (39b), it will attempt to combine them through Minimal Generalisation – that is, as conservatively as possible. The minimal generalisation for (39a) and (39b) is (39c), which retains as much as possible of the original two rules. However, in the implementation of MGL from Albright & Hayes (2002, 2003), when two rules are combined, the longest substrings shared by the rules are retained – in this case /p/ – but segment combination (e.g., {ʊ,o}) proceeds only one position further; everything else is replaced by a free variable X (see Albright & Hayes 2002: 60 for a complete description of this process). Thus, their implementation returns (39d), which is less conservative than the actual minimal generalisation (39c).

(39) a. y → u / bʊp __ #
b. y → u / dop __ #
c. y → u / {b, d}{ʊ, o}p __ #
d. y → u / X{ʊ, o}p __ #

This issue does not arise in the original articles by Albright & Hayes (2002, 2003), because the object of study was the English past tense, in which the surface allomorph is determined by an immediately adjacent segment. Thus, any regularities beyond the adjacent segment, which their implementation would miss, would be spurious anyway. However, for the purposes of this experiment, the implementation is problematic. Consequently, I used my own implementation of MGL, which correctly generates the minimally general combination of rules.[13] I use the rule with the minimally general context in which /-y/ surfaces as [-u] to produce a surface form for each test instance.

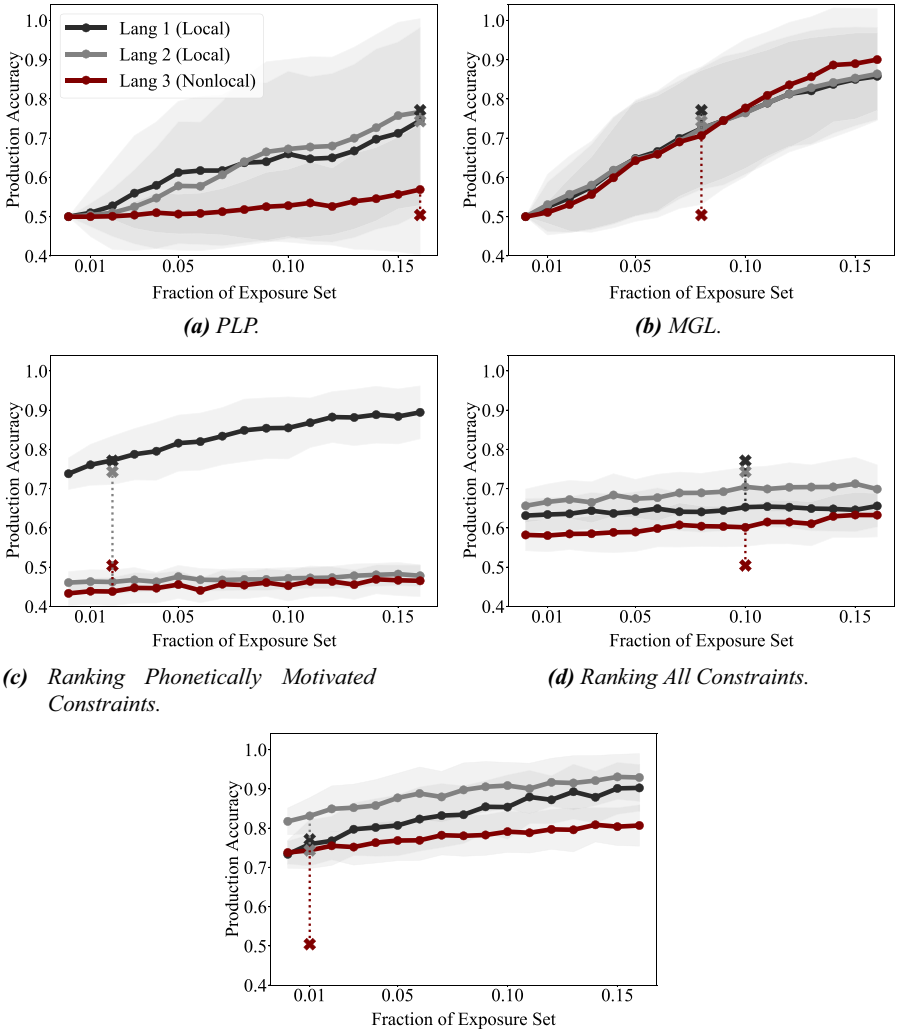
*4.2.2. Results*

Figure 2 shows the results. The *x*-axis of each plot is the free variable discussed above, measuring the fraction of the exposure set that the learner successfully constructed a UR–SR pair for. The *y*-axis reports the average model performance (over the 100 simulations) for each language. The points marked with a colour-coded × provide the average performance over the 20 human participants from Baer-Henney & van de Vijver (2012). Since each model gets to optimise over the free variable, I select, for each respective model, the *x*-value where it best matches the human performance, averaged over all three languages. This point can be seen by where the × marks are placed. A colour-coded vertical line from each human performance marker to the corresponding model performance shows the difference between the two.

[12] Baer-Henney & van de Vijver (2012) used both high- and low-frequency settings, where the high-frequency setting included a higher fraction of plural forms in the exposure set. Since I already treat the amount of exposure data available for learning phonology as a free variable, I followed the high-frequency setting for the experiment.

[13] All other experiments involving MGL used Albright & Hayes's (2002, 2003) implementation.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press




***(a)*** *PLP.*

***(b)*** *MGL.*

***(c)*** *Ranking Phonetically Motivated Constraints.*

***(d)*** *Ranking All Constraints.*

***(e)*** *Ranking Language-Relevant Constraints.*

***Figure 2.*** *Plots show model accuracies on test words when trained on each of the artificial languages from Baer-Henney & van de Vijver (2012), as a function of the fraction of the exposure set they are trained on. The × marks show human generalisation accuracy, at the x-axis point where each model best matches human generalisation behaviour. PLP and MGL's results are in (a) and (b), while grammars learned by ranking a provided constraint set are in (c)–(e). In (c), the ranked constraints are all phonetically motivated; in (d), all constraints needed for the three languages are included and in (e), each model for each language ranks only the constraints relevant to that language.*

PLP (Figure 2a) is the best match to the human results, learning the more local generalisations (37a) and (37b) substantially more easily than the less local (37c). This is because PLP requires sufficient evidence against local generalisations (per the Tolerance Principle) before it will abandon them for less local ones (§2.2.2). This is reminiscent of Gómez & Maye's 2005: 199 characterisation of human learners as attending to local contexts even 'past the point that such structure is useful' before eventually moving on to less local information. In contrast, MGL (Figure 2b) learns all three generalisations equally well because it constructs the most conservative – and hence widest-context – generalisation that is sustainable. If a grammar is constructed by ranking a universal constraint set that includes only phonetically motivated constraints (Figure 2c), only the generalisation in (37a) can be learned because

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

that is the only phonetically motivated generalisation. On the other hand, if all relevant constraints are included together (Figure 2d) or on their own (Figure 2e), all three generalisations are learned roughly equally well. This is because learning reduces to constraint ranking – the constraints being provided – and thus fails to distinguish between more and less local constraints. In PLP's learning of the pattern in (37c), the number of exceptions to the more local generalisation will eventually become too numerous under the Tolerance Principle, and PLP will construct a less local rule, which will correctly characterise the alternation. Thus, PLP predicts that given sufficient time and data, learners will eventually be able to learn the alternation in (37c).

#### *4.2.3. Takeaways*

PLP reflects human learners' preference for local generalisations (**Q3**).

### *4.3. Learning German devoicing*

I now evaluate PLP on syllable-final obstruent devoicing in German (Wiese 1996).

#### *4.3.1. Setup*

This experiment simulates child acquisition by using vocabulary and frequency estimations from child-directed speech in the Leo corpus (Behrens 2006). I retrieved the corpus from the CHILDES database (MacWhinney 2000) and intersected the extracted vocabulary with CELEX (Baayen *et al.* 1996) to get phonological and orthographic transcriptions for each word. I also computed the frequency of each word in the Leo corpus. The resulting dataset consists of 9,539 words. To construct URs and SRs, I followed Gildea & Jurafsky (1996), using the CELEX phonological representations as SRs and discrepancies between CELEX phonology and orthography to construct URs, since German orthography does not reflect devoicing. Specifically, I make the syllable-final obstruents voiced for the URs of all words where the corresponding orthography indicates a voiced obstruent. In this set of data, only 8.2% of words involve devoicing, which means a substantial number of SRs are unchanged by this process from the corresponding URs. However, this is an appropriate and realistic scenario, since the data were constructed from child-directed speech, and is thus a reasonable approximation of the data that children have access to when learning this generalisation.

The experimental procedure samples one word at a time from the data, weighted by frequency. The word is presented to each model and added to its vocabulary. Sampling is with replacement, so the learners are expected to encounter the same word multiple times, at frequencies approximating what a child would encounter. When the vocabulary reaches a size of 100, 200, 300 and 400, each model is probed to produce an SR for each UR in the dataset that is *not* in the vocabulary (i.e., held-out test data). The fraction of these predictions that it gets correct is reported as the model's accuracy. The models MGL, ED and OSTIA are designed as batch learners, so they are trained from scratch on the vocabulary before each evaluation period.[14] PLP, UCON and ORACLE learn incrementally.

This simulation is carried out 10 times to simulate multiple learning trajectories. The results are averages and standard deviations over these 10 runs.

ORACLE is provided with the constraints in (40).

(40) MAX, DEP, IDENT(VOICE), IDENT(SON), IDENT(NAS), *[+voi, −son] $]_\sigma$

The markedness constraint *[+voi, −son] $]_\sigma$, which marks syllable-final voiced obstruents, is the relevant markedness constraint for this process. UCON is supplied with two additional constraints: *NÇ, which marks voiceless consonants following nasals, and *COMPLEX. Both are frequently considered to be universal, violable constraints (Locke 1983; Rosenthall 1989; Prince & Smolensky 1993; Pater 1999). I included these to capture the assumption of a universal constraint set, which requires learning that *COMPLEX and *NÇ are violable in German; for instance, /glaubənd/ → [glau.bən̲t̲] ('believing') violates *COMPLEX and *NÇ.

[14] I provide MGL the frequency with which each vocabulary word has appeared, which it can make use of.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

***Table 1.*** *Model accuracies (with standard deviations) on held-out test data at different training vocabulary sizes. PLP readily learns an accurate generalisation for German syllable-final obstruent devoicing*

| Model | Vocabulary size | | | |
|---|---|---|---|---|
| | 100 | 200 | 300 | 400 |
| PLP | **1.000 ± 0.00** | **1.000 ± 0.00** | **1.000 ± 0.00** | **1.000 ± 0.00** |
| MGL | 0.918 ± 0.00 | 0.918 ± 0.00 | 0.918 ± 0.00 | 0.919 ± 0.00 |
| ED | 0.008 ± 0.00 | 0.178 ± 0.03 | 0.389 ± 0.04 | 0.543 ± 0.04 |
| OSTIA | 0.023 ± 0.02 | 0.022 ± 0.01 | 0.031 ± 0.01 | 0.040 ± 0.00 |
| UCON | 0.960 ± 0.03 | 0.988 ± 0.00 | 0.992 ± 0.00 | 0.995 ± 0.00 |
| ORACLE | 0.982 ± 0.01 | 0.997 ± 0.00 | 0.998 ± 0.00 | 0.999 ± 0.00 |
| ID | 0.918 ± 0.00 | 0.918 ± 0.00 | 0.918 ± 0.00 | 0.918 ± 0.00 |

*4.3.2. Results*

The results are shown in Table 1. PLP learns an accurate grammar, which consists of the single generalisaton shown in (41).

(41) $[+\text{voi}, -\text{son}] \rightarrow [-\text{voi}] \; / \; \_\_ \; ]_\sigma$

While PLP achieves perfect accuracy by the time the vocabulary has grown to size 100, it does produce errors in the process of getting there. A primary example is underextensions. In my experiments, underlyingly voiced stops tended to enter the vocabulary earlier than voiced fricatives. Consequently, PLP sometimes fails to extend devoicing to fricatives until evidence of them devoicing enters the vocabulary. These underextensions are over held-out test words – that is, words not in the learner's vocabulary. Thus, this is a prediction about an early state of the learner's phonological grammar, and not a prediction that children go through a stage of voicing final voiced fricatves. Indeed, I found that as soon as an instance of fricative devoicing enters the vocabulary, PLP extends the generalisation to account for it.

Ranking a provided constraint set (ORACLE and UCON) can yield the same generalisation as PLP: the sequence $[+\text{voi}, -\text{son}]]_\sigma$ is not allowed in German, and violations of this restriction are repaired by devoicing. But the differences in how PLP learns this generalisation are informative. Both UCON and ORACLE are provided with the knowledge that the sequence $[+\text{voi}, -\text{son}]\ ]_\sigma$ is marked. In contrast, PLP discovers the marked sequence in the process of learning.

In German, the onset [bl] is allowed (e.g., /blɑu/ → [blɑu]). PLP always produces the correct SR for /blɑu/ as a consequence of its identity default (Table 2). Whether a constraint-ranking model incorporates a preference for identity between inputs and outputs depends on what constraints it ranks. Because ORACLE ranks only the constraints active in the language being learned, it – like PLP– does not produce unmotivated errors. If a universal constraint set is ranked (UCON), then markedness constraints that are violable in the language being learned will lead to unmotivated errors. For instance, prior to downranking *COMPLEX, UCON sometimes produces [bəlɑu] for /blɑu/, with the sequence /bl/ broken up by an epenthetic schwa, even though complex onsets are allowed in German. However, deletion tends to be more common than epenthesis as a repair in child utterances, and it appears to be due to articulatory limitations rather than to the child's hypothesised adult grammar.

Both UCON and ORACLE sometimes produce /kɪnd/ as *[kɪndə] or *[kɪn] rather than [kɪnt], because they must figure out the relative ranking of faithfulness constraints in order to capture which repair German uses to avoid violations of $*[+\text{voi}, -\text{son}]\ ]_\sigma$. In contrast, PLP infers the repair – devoicing – directly from the discrepancy it observes in the data.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

***Table 2.*** *Analysis of the types of errors each of the models that learn an accurate grammar makes in the process. Because it adds generalisations to the grammar only when necessitated by surface alternations, PLP produces no unmotivated errors*

| Error type | Example | PLP | UCON | ORACLE |
|---|---|---|---|---|
| Unmotivated | /blɑu/ → *[bə̲lɑu] | no | yes | no |
| Wrong repair | /kɪnd/ → *[kɪnd̲ə] | no | yes | yes |
| Under- or overextension | /kɪnd/ → *[kɪn̲d] | yes | yes | yes |

None of the other models perform competitively: PLP outperforms them all by a statistically significant amount ($p < 0.01$), as measured by a paired $t$-test against the null hypothesis that each model's performance over the 10 simulations has the same average accuracy as PLP's. MGL, which generalises as conservatively as possible, struggles to generalise beyond the training data. This is seen in its slow rate of improvement. ED is a powerful model in natural language processing when substantial amounts of data are available, but it struggles to learn on the small vocabularies at the scale children learn from. OSTIA struggles even more, consistent with the negative results of Gildea & Jurafsky (1996), who presented it with much larger vocabularies.

#### *4.3.3. Takeaways*

PLP is readily able to learn German syllable-final devoicing (**Q2**) and never introduces unmotivated generalisations (**Q3**).

#### *4.3.4. Opacity*

The associate editor observed that devoicing in Polish interacts opaquely with o-raising, in which /ɔ/ surfaces as [u] before word-final oral consonants that are underlyingly voiced (Kenstowicz 1994; Sanders 2003). As a proof of concept, I ran PLP on the data in Sanders (2003: 48–51). PLP learned the counterbleeding rule system in (42), with raising ($r_1$) correctly ordered before devoicing ($r_2$).[15]

(42) $G = r_2 \circ r_1$, where

$r_1 = ɔ \rightarrow u \;/\; \_\_\; [+voi]\,\#$
$r_2 = [+voi, -son] \rightarrow [-voi] \;/\; \_\_\; \#$

Rule $r_2$ accounts for devoicing both in isolation (43a) and in words exhibiting raising (43c). Rule $r_1$ accounts for raising both in isolation (43b) and when its underlying context is opaquely obscured by devoicing (43c).

(43) a. /klub/ → [klup] 'club' NOM.SG
b. /bɔl/ → [bul] 'ache' NOM.SG
c. /bɔb/ → [bup] 'bean' NOM.SG

The correct ordering was achieved because the reverse ordering, in which devoicing bleeds raising, results in errors like *[bɔp] for /bɔb/. This demonstrates that PLP is capable of handling at least this case of opacity. I leave a systematic study of opacity for future work (see §2.3.3).

[15]The examples from Sanders (2003) were too sparse to distinguish between [+voi]# and [+cons,+voi,−nas]# as the context for raising; a more realistic lexicon should drive PLP to the more nuanced context.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

### *4.4. Learning a multi-process grammar*

This experiment evaluates PLP at learning multiple generalisations simultaneously. The processes modelled are the English alternating plural and 3SG.PRES affix /-z/ (44a), the alternating past tense affix /-d/ (44b), and vowel nasalisation (44c).

(44) a. /dɑg-z/ → [dɑgz]
/wɔk-z/ → [wɔks]
/hɔrs-z/ → [hɔrsəz]
b. /smɛl-d/ → [smɛld]
/wɔk-d/ → [wɔkt]
/foʊld-d/ → [foʊldəd]
c. /ðɛm/ → [ðɛ̃m]
/sʌmθɪŋ/ → [sʌ̃mθɪ̃ŋ]
/dæns/ → [dæ̃ns]

#### *4.4.1. Setup*

This experiment, like the first, simulates child language acquisition. The child-directed speech is aggregated across English corpora in CHILDES (MacWhinney 2000), including the frequency of each word. Only words with '%mor' tags were retained, because the morphological information was needed to construct URs. Transcriptions from the *CMU Pronouncing Dictionary* (Weide 2014) served as SRs, with nasalisation added to vowels preceding nasal consonants. URs had all vowels recorded without nasalisation. The URs of the affixes on all past tense verbs, plural nouns and 3SG.PRES verbs were set to /d/, /z/ and /z/, respectively. The resulting dataset contains 20,421 UR–SR pairs.

The experimental procedure is the same as for German, sampling words weighted by frequency and reporting accuracies at predicting SRs from URs over held-out test words when each learner's vocabulary reaches certain sizes: 1K, 2K, 3K and 4K words.

I omit results from MGL, ED and OSTIA because they continued to be non-competitive. ORACLE once again ranks only the relevant constraints, listed in (45), and UCON receives these plus *COMPLEX and *NÇ.

(45) CON = {
MAX, DEP, IDENT(VOICE), IDENT(SON), IDENT(NAS),
AGREE(VOICE), *SS, *[+vowel, −nas] [+cons, +nas],
*[−cont, −dist, −son] [−cont, −dist, −son]
}

All faithfulness constraints other than DEP were split into two – one for stems and one for affixes – so that, for instance, *[wɔgz] could be ruled out for input /wɔk-z/ in (44a).

#### *4.4.2. Results*

The models' accuracies on held-out test words, shown in Table 3, reveal that PLP learns an accurate grammar by the time its vocabulary grows to about 2,000 words. PLP's output is shown as an ordered list of rules in (46).

(46) $G = r_5 \circ r_4 \circ r_3 \circ r_2 \circ r_1$, where

$$r_1 = [+\text{syl}] \rightarrow [+\text{nas}] \,/\, \_\_\, [+\text{nas}]$$
$$r_2 = \varnothing \rightarrow \text{ə} \,/\, [+\text{sib}] \,\_\_\, [+\text{sib}]$$
$$r_3 = [+\text{sib}, +\text{voi}] \rightarrow [-\text{voi}] \,/\, [-\text{voi}] \,\_\_$$
$$r_4 = \varnothing \rightarrow \text{ə} \,/\, [+\text{cor}, -\text{cont}, -\text{nas}] \,\_\_\, [+\text{cor}, -\text{cont}, -\text{nas}]$$
$$r_5 = [+\text{cor}, -\text{cont}, -\text{nas}, +\text{voi}] \rightarrow [-\text{voi}] \,/\, [-\text{voi}] \,\_\_$$

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

***Table 3.*** *Model accuracies (with standard deviations) on held-out test data at different training vocabulary sizes. PLP readily learns an accurate grammar for the English processes in* (44)

| Model | Vocabulary size | | | |
|---|---|---|---|---|
| | 1,000 | 2,000 | 3,000 | 4,000 |
| PLP | **0.984 ± 0.01** | **0.992 ± 0.00** | **0.995 ± 0.00** | **0.997 ± 0.00** |
| UCON | 0.969 ± 0.00 | 0.982 ± 0.00 | 0.987 ± 0.00 | 0.990 ± 0.00 |
| ORACLE | 0.980 ± 0.00 | 0.989 ± 0.00 | 0.991 ± 0.00 | 0.992 ± 0.00 |
| ID | 0.510 ± 0.00 | 0.510 ± 0.00 | 0.510 ± 0.00 | 0.510 ± 0.00 |

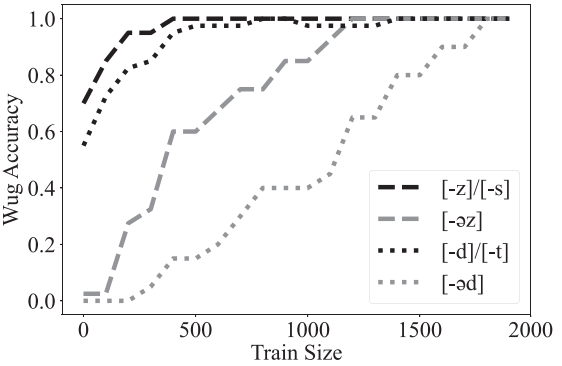



***Figure 3.*** *PLP's accuracy on the plural and past tense nonce words from Berko (1958) as training progressed. The black dashed line denotes plurals that should take [-z] or [-s] and the grey dashed lines those that should take [-əz]. The dotted lines represent the analogues for past tense. The fact that [-z]/[-s] accuracy converges before [-əz] and [-d]/[-t] before [-əd] matches Berko's finding that children learn [-z]/[-s] and [-d]/[-t] before [-əz] and [-əd].*

The rules were ordered as described in §2.3.3, with $r_2$ before $r_3$ and $r_4$ before $r_5$ (bleeding order) being the inferred ordering dependencies. Thus, as described in §2.3.3, PLP learned that epenthesis bleeds devoicing. Rules $r_2$–$r_5$ do not encode a word-final context because they satisfy the Tolerance Principle threshold without it, so there is no need for PLP to expand the search window. The extension of [+cor, −cont, −nas] is {t, d}, and the extension of [+cor, −cont, −nas, +voi] is {d}.

The fact that no model achieves 100% accuracy is due to a handful of words that do not follow the generalisations in (44). For instance, compounds like [bɛdtaĩm] allow the sequence [dt], but the models predict there should be an epenthetic vowel to split the sequence. Such exceptions are easily accounted for if I assume the learner recognises the word as a compound. Since exceptions are inevitable in naturalistic data, I chose to not remove them.

Berko's (1958) seminal study found that children aged 4–7 years could accurately inflect nonce words that take the [-z], [-s], [-d] or [-t] suffixes, but that they performed much worse at inflecting nonce words taking the [-əz] or [-əd] suffixes. Adults could inflect nonce words with [-əz] or [-əd], suggesting that voicing assimilation process may be learned earlier than the epenthesis process. I show PLP's accuracy on Berko's different categories of nonce words in Figure 3 as the vocabulary grows ($x$-axis). PLP's accuracy on nonce words taking [-z] or [-s] (dashed black line) converges earlier than its accuracy on nonce words taking [-əz] (dashed grey line); similarly the accuracy for nonce words taking [-d] or [-t] (dotted black line) converges earlier than for nonce words taking [-əd] (dotted grey line). Thus, the order of acquisition matches Berko's finding.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

#### *4.4.3. Takeaways*

The results in this more challenging setting, where multiple processes are simultaneously active, support the takeaways from the prior experiment. PLP successfully learns all the generalisations (**Q2**) and does not introduce unmotivated generalisations (**Q3**).

### *4.5. Learning Tswana's post-nasal devoicing*

Although a majority of phonological patterns may be phonetically grounded, some processes nevertheless appear to lack or even oppose phonetic motivation (Anderson 1981; Buckley 2000; Johnsen 2012; Beguš 2019). Moreover, these must still be learnable, because children continue to successfully acquire them (Johnsen 2012: 506). An example of such a pattern is the post-nasal devoicing in Tswana shown above in (2), which Coetzee & Pretorius (2010) confirmed to be productive despite operating against the phonetic motivation for post-nasal voicing (Hayes & Stivers 2000; Beguš 2019). Beguš (2019: 699) found post-nasal devoicing reported as a sound change in 13 languages and dialects, from eight language families.

Models of phonological learning should account for the fact that productive patterns are successfully learned by humans even if they are not phonetically grounded. A consequence of PLP's identity default is that generalisations are added to the grammar whenever they are motivated by surface alternations. Since surface alternations in Tswana motivate a generalisation for post-nasal devoicing, PLP should be able to acquire the Tswana pattern. This experiment attempts to confirm this (**Q3**).

#### *4.5.1. Setup*

For this experiment, I used the 10 UR–SR pairs from Coetzee & Pretorius (2010: 406) as training data. Five pairs involve devoicing resulting from the ISG.OBJ clitic /m/ attaching to a stem that starts with a voiced obstruent. The other five pairs involve the IPL.OBJ clitic /re/ attaching to the same stems, which serve as negative examples since the clitic does not introduce a nasal. These data are not necessarily representative of the data a child would have during acquisition, and the learnability experiment thus serves only as a proof of concept.

The test data consist of the same 20 /b/-initial nonce words presented to the participants in Coetzee & Pretorius (2010: 407) – 10 stems each combined with /m/ and /re/.

#### *4.5.2. Results*

The results in Table 4 demonstrate that PLP can learn Tswana's post-nasal devoicing without requiring the existence of a phonetically unmotivated universal constraint.[16] Constraint-ranking models can also learn the generalisation but depend on an account of how the constraint *NÇ, which penalises what is

***Table 4.*** *PLP learns precisely the set of processes active in its experience. This provides a straightforward account of how productive phonological processes can be learned even if they operate against apparent phonetic motivation, like devoicing in Tswana following nasals (Coetzee & Pretorius 2010). With PLP, the unmotivated constraint *NÇ need not be assumed to be universal*

| Model | Generalisation | Test accuracy |
|---|---|---|
| PLP | b→ [−voi] / m__ | 1.0 |
| Ranking without *NÇ | {*NÇ, IDENT(VOICE)} | 0.5 |
| Ranking with *NÇ | *NÇ ≫{*NÇ, IDENT(VOICE)} | 1.0 |

[16]PLP learns *[mb] rather than *NÇ because the training data only included [mb] instances; if more representative training data were available, PLP would induce natural classes, as in the previous experiments.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

not usually considered to be a universally marked sequence (Locke 1983; Rosenthall 1989; Pater 1999; Beguš 2016, 2019), is added to the constraint set.

#### *4.5.3. Takeaways*

Because PLP assumes UR–SR identity by default, it constructs precisely the generalisations necessary to account for the discrepancies active in its experience, providing a straightforward account of how productive generalisations can be learned even if they are opposed to apparent phonetic motivation, as humans evidently do (**Q3**; see Seidl & Buckley 2005, Johnsen 2012: 506; Beguš 2018, ch. 6).

## 5. Discussion

### *5.1. The nature of locality*

One reviewer asked what sort of tendency I view locality to be. I view the cognitive tendency for humans to prefer constructing local generalisations to be a geometric computational consequence. That is, if words are viewed, at least to a first approximation, as linear objects, this linear geometry introduces the notion of locality as small linear distance. In my view, the reason that a human is more likely to construct a generalisation that conditions $x_i$ on $x_{i-1}$ than on $x_{i-2}$ in a sequence $\ldots, x_{i-2}, x_{i-1}, x_i$ (see §1.1) is that a search outward from $x_i$ encounters $x_{i-1}$ before it encounters $x_{i-2}$.

PLP is an attempt to state this in explicit computational terms. An immediate consequence of this hypothesis is that if $x_{i-1}$ is sufficient to account for whatever the uncertainty in $x_i$ is (e.g., what its surface form is), then $x_{i-2}$ will never be considered, even if there is some statistical dependency between the two. I believe this prediction is consistent with the experimental results from sequence learning, discussed in §1.1, in which participants tracked adjacent dependencies even when non-adjacent dependencies were more statistically informative (Gómez & Maye 2005) and constructed local generalisations rather than less local ones when the exposure data underdetermined the two (as in the poverty-of-stimulus paradigms of Finley 2011 and McMullin & Hansson 2019). I note that words may not be *exactly* linear – segment articulations have gestural overlap, syllables are often viewed as hierarchical structures, and representations like autosegmental tiers may be present. However, I think treating words as linear sequences is a good first approximation. Work on tier-locality also recognises that string-locality is a special case of tier-locality in which all segments are on the same tier (see, e.g., Hayes & Wilson 2008; Heinz *et al.* 2011; McMullin 2016).

An alternative view could be that locality is distributional: a learner may track the dependency between $x_i$ and both $x_{i-1}$ and $x_{i-2}$, and may find that $x_{i-1}$ is more statistically robust as a generalisation, preferring it for that reason. However, this view is inconsistent with the findings that when statistical robustness is controlled (Finley 2011; McMullin & Hansson 2019), and even when it favours the less local dependency (Gómez & Maye 2005), humans systematically generalise locally. The distributional approach could be combined with a stipulated bias (prior) favouring local dependencies, but this would simply describe the phenomenon, not explain it.

### *5.2. Future directions*

Research on phonological representations recognises that strict locality arises not only over string representations but also over representations like tiers and metrical grids (Goldsmith 1976; Hayes & Wilson 2008; Heinz *et al.* 2011; McMullin 2016). PLP could be extended to construct generalisations over these representations. However, an account of how a learner may construct increasingly abstract representations should first be given. Recent work has investigated this, proposing an abductive algorithm in which learners iteratively propose new representations in response to alternations not being sufficiently predictable from adjacent dependencies in a linear representation (Belth in press). This mirrors PLP's iterative expansion of an attention window. Future work will investigate combining these learning models: it remains an open question how the mechanisms of expanding attention and projecting new representations interact.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

In another direction, variation is an important aspect of phonology (Coetzee & Pater 2011). PLP currently learns categorical processes, and future work will investigate extending PLP to handle variation by allowing rules to be probabilistic.

## A. PLP and strict locality

In this appendix, I discuss how PLP's generalisations can be characterised in the formal-language-theoretic terms of strict locality. §A.1 shows that the sequences PLP learns are strictly local definitions, and thus have the interpretation of banning substrings (Heinz 2018: 28). In §A.2, I then discuss how PLP's generalisations describe input-strictly local maps.

### *A.1. Strict locality of sequences*

Strictly local stringsets (McNaughton & Papert 1971) are stringsets whose members 'are distinguished from non-members purely on the basis of their $k$-factors' (Rogers *et al.* 2013: 98). A $k$-factor of a string is a length-$k$ substring, and the set of $k$-factors over an alphabet $\Sigma$ is $F_k(\Sigma^*) = \{w \in \Sigma^* : |w| \leq k\}$ (Rogers *et al.* 2013: 96). A *strictly k-local definition* $\mathcal{G}$ is a subset of the $k$-factors over $\Sigma$, that is, $\mathcal{G} \subseteq F_k(\Sigma^*)$.[17] A definition is a strictly local definition if it is strictly $k$-local for some $k$. To be shown here is that the sequences PLP learns, as defined in (6), repeated in (47), are strictly local definitions.

(47) $\mathcal{S} \triangleq \bigcup_{k=1}^{\infty} \{s_1 s_2 \ldots s_k : s_i \subset \Sigma\}$

Since $\bar{s} \in \mathcal{S}$ is a sequence of sets of segments $s_i \subset \Sigma$, we can define the *extension*, $E_{\bar{s}}$, of $\bar{s}$ as the set of sequences of segments that match $\bar{s}$, as in (48), where $k = |\bar{s}|$.

(48) $E_{\bar{s}} \triangleq \{a_1 a_2 \ldots a_k : a_i \in s_i \; \forall i \in 1 \ldots k\}$

For example, a sequence of two adjacent sibilants (49a) has the extension (49b).

(49) a. $\bar{s}$ = [+sib][+sib]

b. $E_{\bar{s}}$ = {ss, sz, zs, ʃz, ʒs, ...}

**Theorem 1.** The instances $E_{\bar{s}}$ of any $\bar{s} \in \mathcal{S}$ form a strictly local definition over the alphabet $\Sigma$.

*Proof.* For any $a_1 a_2 \ldots a_k \in E_{\bar{s}}$, each $a_i$ is an element of $s_i$ (i.e., $a_i \in s_i$) by (48) and thus an element of $\Sigma$ (i.e., $a_i \in s_i \subset \Sigma$) by (47). Thus, every $a_1 a_2 \ldots a_k \in E_{\bar{s}}$ is a length-$k$ string from $\Sigma^*$. It follows that $E_{\bar{s}} \subseteq F_k(\Sigma^*)$ and that, for $k = |\bar{s}|$, $E_{\bar{s}}$ is a strictly local definition.

### *A.2. Strict locality of generalisations*

Chandlee *et al.* (2014: 40) provides formal-language-theoretic and automata-theoretic definitions of *input strictly local* string-to-string functions, which, for input and output alphabets $\Sigma$ and $\Gamma$, have the following interpretation:

**Definition 1 ($k$ISL function - Informal).** A function (map) $f : \Sigma^* \rightarrow \Gamma^*$ is input strictly local (ISL) iff $\exists k \in \mathbb{N}$ such that each output symbol $o \in \Gamma$ is determined by a length-$k$ window around its corresponding input symbol.[18]

Each of PLP's generalisations is interpretable as a rule of the form (50) with a target context (*cad*) of finite length $|cad|$,[19] and under simultaneous application (cf. §2.3.4).

(50) $a \rightarrow b \,/\, c \_\_ \; d$

[17] Rogers *et al.* (2013) add word-initial and word-final markers ($\rtimes$ and $\ltimes$) to $\Sigma$. I assume the learner's segment inventory already contains symbols for syllable and word boundaries.

[18] Length $k$ includes the corresponding input symbol.

[19] Under the realistic assumption that input strings are of finite length.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

Chandlee *et al.* (2014: 41) provides an algorithm for constructing, from any such rule (i.e., with finite target context and under simultaneous application), a finite-state transducer with the necessary and sufficient automata-theoretic properties of an ISL map. Consequently, if Chandlee's algorithm is a valid constructive proof, it follows that each generalisation that PLP constructs describes an ISL map. When these are combined into a grammar, it is unknown whether the resulting grammar is also ISL, because it is unknown whether ISL maps are closed under composition (Chandlee 2014: 149).

## B. Differences between PLP and MGL

PLP differs in several ways from the MGL model of Albright & Hayes (2002, 2003). Note that PLP is designed to learn phonology, while MGL was designed for producing English past-tense inflections from verb stems, though it can be extended to other settings.

### *B.1. Generalisation strategy*

PLP and MGL use different generalisation strategies: PLP generalises as *locally* as possible, and MGL generalises as *conservatively* as possible. As discussed in §1.1 and tested in §4.2, I believe that PLP's generalisation strategy is better supported by studies of human learning.

### *B.2. Number of rules*

Another difference between PLP and MGL is the number of rules they generate. For German syllable-final devoicing at a vocabulary size of 400 (§4.3), PLP learns the single rule in (51).

(51) $[+\text{voi},-\text{son}] \rightarrow [-\text{voi}] \ / \ \_\_ \ ]_{\sigma}$

In contrast, MGL learns 102 rules for where devoicing should take place and 4,138 for where it should not. An example of the former is (52a) and the latter (52b) (both are presented with the extensions of the natural classes for clarity). Rules like (52b) are learned because not every word involves devoicing, and thus MGL needs such rules in order to produce those words (§B.3).

(52) a. $\text{g} \rightarrow \text{k} \ / \ \{\text{a,e,i,o,u,y,ø,æ,ɔ,ə,ɛ,ɪ,ʊ}\} \ \_\_ \ ]_{\sigma} \ \#$

⋮

b. $\varnothing \rightarrow \varnothing \ / \ \{\text{f,k,p,t,x}\}]_{\sigma} \ \#$

⋮

### *B.3. Production*

MGL may produce multiple candidate outputs for an input, because every rule that applies to the input generates a candidate output. The quality of a candidate output is 'the confidence of the best rule that derives it' (Albright & Hayes 2002, §3.2). I used the candidate with the highest confidence as MGL's prediction. This differs from PLP's production (§2.3.4), which applies all rules (here just one) in order. This difference is not significant when learning a single phonological process, but it is not straightforward to use MGL to learn multiple processes simultaneously. For instance, in §4.4, for input /ɪnsɛkt-z/, MGL's rule(s) for vowel nasalisation may produce the candidate *[ĩnsɛktz], and its rule(s) for pluralisation may produce the candidate *[ɪnsɛkts]. However, MGL does not provide a mechanism to apply *both* rules to produce the correct output [ĩnsɛkts].

### *B.4. Natural classes*

MGL's natural-class induction differs from PLP's in two ways. First, MGL does not form natural classes for every part of a rule. For example, the two rules in (53a) will combine to form a third (53b) – and

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

similarly for (53c) and (53d) – but rules (53b) and (53d) will not combine to form (53e), because only contexts (not targets) are merged.

(53) a. ə → ə̃ / __ n
ə → ə̃ / __ m
b. ə → ə̃ / __ {n, m}
c. ʌ → ʌ̃ / __ n
ʌ → ʌ̃ / __ m
d. ʌ → ʌ̃ / __ {n, m}
e. {ə, ʌ} → [+nas] / __ {n, m} (formed by PLP but not by MGL)

Moreover, when rules are combined, the new rule and the original rules are all retained. In contrast, PLP will construct (53e), and only it will be present in the grammar (§2.3.1).

Second, when MGL creates natural classes for a set of segments, it retains all features shared by those segments, whereas PLP only retains those needed to keep the rule satisfactorily accurate. Thus, for (54a) MGL will construct (54b), while PLP will construct (54c).

(54) a. ə → ə̃ / __ {n, m}
b. ə → ə̃ / __ [+ant, +cons, +lab, +nas, +son, +voi, −back, −cg, −cont, −cor, −delrel, −hi, −lat, −lo, −long, −round, −sg, −syl, −velaric]
c. ə → ə̃ / __ [+nas]

Consequently, PLP will correctly extend ə-nasalisation to contexts with a following ŋ, but MGL will need to wait for such an instance in the training data before constructing the full generalisation.

**Acknowledgements.** This work has been greatly improved through the thoughtful and constructive comments of three anonymous reviewers, and the associate editor. I thank Andries Coetzee, Charles Yang, Jane Chandlee and Jeffrey Heinz for many discussions throughout the development of this project. This work was supported by an NSF GRF. All mistakes are my own.

**Competing interests.** The author declares no competing interests.

## References

Aksënova, Alëna (2020). SigmaPie. Available at https://github.com/alenaks/SigmaPie.

Albright, Adam (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology* **26**, 9–41.

Albright, Adam & Bruce Hayes (2002). Modeling English past tense intuitions with minimal generalization. In *Morphological and phonological learning: proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*. Philadelphia: Association for Computational Linguistics, 58–69.

Albright, Adam & Bruce Hayes (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* **90**, 119–161.

Anderson, Stephen R. (1981). Why phonology isn't 'natural'. *LI* **12**, 493–539.

Aslin, Richard N., Jenny R. Saffran & Elissa L. Newport (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* **9**, 321–324.

Baayen, R. Harald, Yu-Ying Chuang & James P. Blevins (2018). Inflectional morphology with linear mappings. *The Mental Lexicon* **13**, 230–268.

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins (2019). The discriminative lexicon: a unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity* **2019**, Article No. 4895891, 39 pp.

Baayen, R. Harald, Richard Piepenbrock & Léon Gulikers (1996). CELEX2. Corpus published online by the Linguistic Data Consortium at https://doi.org/10.35111/gs6s-gm48.

Baer-Henney, Dinah & Ruben van de Vijver (2012). On the role of substance, locality, and amount of exposure in the acquisition of morphophonemic alternations. *Laboratory Phonology* **3**, 221–249.

Baković, Eric (2011). Opacity and ordering. In John Goldsmith, Jason Riggle & Alan C. L. Yu (eds.) *The handbook of phonological theory*, 2nd edition. Oxford: Wiley-Blackwell, 40–67.

Beguš, Gašper (2016). Post-nasal devoicing and a probabilistic model of phonological typology. Ms, Harvard University.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

Beguš, Gašper (2018). *Unnatural phonology: a synchrony–diachrony interface approach*. PhD dissertation, Harvard University.
Beguš, Gašper (2019). Post-nasal devoicing and the blurring process. *JL* **55**, 689–753.
Beguš, Gašper (2022). Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks. *Computer Speech & Language* **71**, Article No. 101244.
Behrens, Heike (2006). The input–output relationship in first language acquisition. *Language and Cognitive Processes* **21**, 2–24.
Belth, Caleb (in press). A learning-based account of phonological tiers. *LI*, 37 pp. Published online March 2024; forthcoming in print.
Belth, Caleb, Sarah Payne, Deniz Beser, Jordan Kodner & Charles Yang (2021). The greedy and recursive search for morphological productivity. *Proceedings of the Annual Meeting of the Cognitive Science Society* **43**, 2869–2875.
Berent, Iris (2013). The phonological mind. *Trends in Cognitive Sciences* **17**, 319–327.
Berent, Iris & Tracy Lennertz (2007). What we know about what we have never heard before: beyond phonetics. *Cognition* **104**, 638–643.
Berko, Jean (1958). The child's learning of English morphology. *Word* **14**, 150–177.
Berkson, Kelly, Stuart Davis & Alyssa Strickler (2017). What does incipient /ay/-raising look like? A response to Josef Fruehwald. *Lg* **93**, e181–e191.
Boersma, Paul (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **21**, 43–58.
Boersma, Paul (1999). Optimality-Theoretic learning in the Praat program. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **23**, 17–35.
Boersma, Paul & Bruce Hayes (2001). Empirical tests of the gradual learning algorithm. *LI* **32**, 45–86.
Boersma, Paul & Joe Pater (2008). *Convergence properties of a gradual learning algorithm for Harmonic Grammar*. Ms, University of Massachusetts Amherst.
Buckler, Helen & Paula Fikkert (2016). Dutch and German 3-year-olds' representations of voicing alternations. *Language and Speech* **59**, 236–265.
Buckley, Eugene (2000). On the naturalness of unnatural rules. *UCSB Working Papers in Linguistics* **9**, 1–14.
Calamaro, Shira & Gaja Jarosz (2015). Learning general phonological rules from distributional information: a computational model. *Cognitive Science* **39**, 647–666.
Chandlee, Jane (2014). *Strictly local phonological processes*. PhD dissertation, University of Delaware.
Chandlee, Jane, Rémi Eyraud & Jeffrey Heinz (2014). Learning strictly local subsequential functions. *Transactions of the Association for Computational Linguistics* **2**, 491–504.
Chandlee, Jane, Jeffrey Heinz & Adam Jardine (2018). Input strictly local opaque maps. *Phonology* **35**, 171–205.
Chomsky, Noam (1959). On certain formal properties of grammars. *Information and Control* **2**, 137–167.
Chomsky, Noam (2005). Three factors in language design. *LI* **36**, 1–22.
Chomsky, Noam & Morris Halle (1965). Some controversial questions in phonological theory. *JL* **1**, 97–138.
Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.
Coetzee, Andries W. (2009). Learning lexical indexation. *Phonology* **26**, 109–145.
Coetzee, Andries W. & Joe Pater (2011). The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan C. L. Yu (eds.) *The handbook of phonological theory*, 2nd edition. Oxford: Wiley-Blackwell, 401–434.
Coetzee, Andries W. & Rigardt Pretorius (2010). Phonetically grounded phonology and sound change: the case of Tswana labial plosives. *JPh* **38**, 404–421.
Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest & Clifford Stein (2009). *Introduction to algorithms*. Cambridge, MA: MIT Press.
Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner & Mans Hulden (2016). The SIGMORPHON 2016 shared task – morphological reinflection. In *Proceedings of the 2016 meeting of SIGMORPHON*. Berlin: Association for Computational Linguistics, 10–22.
Emond, Emeryse & Rushen Shi (2021). Infants' rule generalization is governed by the Tolerance Principle. In Danielle Dionne & Lee-Ann Vidal Covas (eds.) *Proceedings of the 45th annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, 191–204.
Ernestus, Mirjam Theresia Constantia & R. Harald Baayen (2003). Predicting the unpredictable: interpreting neutralized segments in Dutch. *Lg* **79**, 5–38.
Fikkert, Paula (1994). *On the acquisition of prosodic structure*. PhD dissertation, Leiden University.
Finley, Sara (2011). The privileged status of locality in consonant harmony. *Journal of Memory and Language* **65**, 74–83.
Finley, Sara & William Badecker (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language* **61**, 423–437.
Fiser, József & Richard N. Aslin (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **28**, 458–467.
Freitas, M. João (2003). The acquisition of Onset clusters in European Portuguese. *Probus* **15**, 27–46.
Gafos, Adamantios I. (1999). *The articulatory basis of locality in phonology*. New York: Garland.
Gildea, Daniel & Dan Jurafsky (1996). Learning bias and phonological-rule induction. *Computational Linguistics* **22**, 497–530.
Gnanadesikan, Amalia (2004). Markedness and faithfulness constraints in child phonology. In René Kager, Joe Pater & Wim Zonneveld (eds.) *Constraints in phonological acquisition*. Cambridge: Cambridge University Press, 73–108.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

Gold, E. Mark (1967). Language identification in the limit. *Information and Control* **10**, 447–474.
Goldsmith, John (1976). *Autosegmental phonology*. PhD dissertation, Massachusetts Institute of Technology.
Gómez, Rebecca (2002). Variability and detection of invariant structure. *Psychological Science* **13**, 431–436.
Gómez, Rebecca & Jessica Maye (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy* **7**. 183–206.
Grijzenhout, Janet & Sandra Joppen (1998). First steps in the acquisition of German phonology: a case study. *Arbeiten des Sonderforschungsbereichs 282* **110**, 24 pp. ROA #304.
Grijzenhout, Janet & Sandra Joppen-Hellwig (2002). The lack of onsets in German child phonology. In Ingeborg Lasser (ed.) *The process of language acquisition*. Frankfurt: Peter Lang, 319–339.
Hale, Mark & Charles Reiss (2008). *The phonological enterprise*. Oxford: Oxford University Press.
Halle, Morris (1978). Knowledge unlearned and untaught: what speakers know about the sounds of their language. In Morris Halle, Joan Bresnan & George A. Miller (eds.) *Linguistic theory and psychological reality*. Cambridge, MA: MIT Press, 294–303.
Hayes, Bruce & Tanya Stivers (2000). *Postnasal voicing*. Ms, University of California, Los Angeles.
Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI* **39**, 379–440.
Heinz, Jeffrey (2018). The computational nature of phonological generalizations. In Larry M. Hyman & Frans Plank (eds.) *Phonological typology*, number 23 in Phonetics and Phonology. Berlin: De Gruyter Mouton, 126–195.
Heinz, Jeffrey, Chetan Rawal & Herbert G. Tanner (2011). Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: human language technologies*. Portland, OR: Association for Computational Linguistics, 58–64.
Howard, Irwin (1972). *A directional theory of rule application in phonology*. PhD dissertation, Massachusetts Institute of Technology.
Hyman, Larry M. (2018). Why underlying representations? *JL* **54**, 591–610.
Jakobson, Roman & Morris Halle (1956). *Fundamentals of language*. The Hague: Mouton.
Jardine, Adam (2016). Computationally, tone is different. *Phonology* **33**, 247–283.
Jardine, Adam, Jane Chandlee, Rémi Eyraud & Jeffrey Heinz (2014). Very efficient learning of structured classes of subsequential functions from positive data. In Alexander Clark, Makoto Kanazawa & Ryo Yoshinaka (eds.) *The 12th International Conference on Grammatical Inference*, number 34 in Proceedings of Machine Language Research. Kyoto: PMLR, 94–108.
Jarosz, Gaja (2019). Computational modeling of phonological learning. *Annual Review of Linguistics* **5**, 67–90.
Johnsen, Sverre Stausland (2012). A diachronic account of phonological unnaturalness. *Phonology* **29**, 505–531.
Johnson, Mark (1984). A discovery procedure for certain phonological rules. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting of the Association for Computational Linguistics*. Stanford, CA: Association for Computational Linguistics, 344–347.
Joos, Martin (1942). A phonological dilemma in Canadian English. *Lg* **18**. 141–144.
Jusczyk, Peter W., Paul Smolensky & Theresa Allocco (2002). How English-learning infants respond to markedness and faithfulness constraints. *Language Acquisition* **10**, 31–73.
Kenstowicz, Michael (1994). *Phonology in generative grammar*. Oxford: Blackwell.
Kenstowicz, Michael & Charles Kisseberth (1979). *Generative phonology: description and theory*. San Diego: Academic Press.
Kerkhoff, Annemarie Odilia (2007). *Acquisition of morpho-phonology: the Dutch voicing alternation*. PhD dissertation, Landelijke Onderzoekschool Taalwetenschap.
Kiparsky, Paul ([1968] 1982). How abstract is phonology? In Paul Kiparsky (ed.) *Explanation in phonology*. Dordrecht: Foris, 119–163. Originally published by the Indiana University Linguistics Club.
Kirov, Christo & Ryan Cotterell (2018). Recurrent neural networks in linguistic theory: revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics* **6**, 651–665.
Koulaguina, Elena & Rushen Shi (2019). Rule generalization from inconsistent input in early infancy. *Language Acquisition* **26**, 416–435.
Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990). *Harmonic Grammar: a formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations*. Technical Report 90-5, Institute of Cognitive Science, University of Colorado, Boulder.
Locke, John L. (1983). *Phonological acquisition and change*. New York: Academic Press.
MacWhinney, Brian (2000). *The CHILDES project: tools for analyzing talk, volume I: transcription format and programs*. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates.
Mayer, Connor (2020). An algorithm for learning phonological classes from distributional similarity. *Phonology* **37**, 91–131.
McCarthy, John J. (2007). Derivations and levels of representation. In Paul de Lacy (ed.) *The Cambridge handbook of phonology*. Cambridge: Cambridge University Press, 99–118.
McCurdy, Kate, Sharon Goldwater & Adam Lopez (2020). Inflecting when there's no majority: limitations of encoder–decoder neural networks as cognitive models for German plurals. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* **58**, 1745–1756.
McMullin, Kevin (2016). *Tier-based locality in long-distance phonotactics: learnability and typology*. PhD dissertation, University of British Columbia.
McMullin, Kevin & Gunnar Ólafur Hansson (2019). Inductive learning of locality relations in segmental phonology. *Laboratory Phonology* **10**. Article No. 14.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press

McNaughton, Robert & Seymour A. Papert (1971). *Counter-free automata*. Cambridge, MA: MIT Press.
Mortensen, David R., Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer & Lori Levin (2016). PanPhon: a resource for mapping IPA segments to articulatory feature vectors. In Yuji Matsumoto & Rashmi Prasad (eds.) *Proceedings of the 26th International Conference on Computational Linguistics: technical papers*. Osaka: COLING, 3475–3484.
Newport, Elissa L. & Richard N. Aslin (2004). Learning at a distance I: statistical learning of non-adjacent dependencies. *Cognitive Psychology* **48**, 127–162.
Oncina, José, Pedro García & Enrique Vidal (1993). Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**, 448–458.
Pater, Joe (1999). Austronesian nasal substitution and other NC effects. In René Kager, Harry van der Hulst & Wim Zonneveld (eds.) *The prosody–morphology interface*. Cambridge: Cambridge University Press, 310–343.
Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal & Emmanuel Dupoux (2006). The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition* **101**, B31–B42.
Prince, Alan & Paul Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Technical Report 2, Rutgers Center for Cognitive Science.
Rasin, Ezer, Iddo Berger, Nur Lan & Roni Katzir (2018). Learning phonological optionality and opacity from distributional evidence. *NELS* **48**, 269–282.
Richter, Caitlin (2018). Learning allophones: what input is necessary. In Anne B. Bertolini & Maxwell J. Kaplan (eds.) *Proceedings of the 42nd annual Boston University Conference on Language Development*, volume 2. Somerville, MA: Cascadilla Press, 659–672.
Richter, Caitlin (2021). *Alternation-sensitive phoneme learning: implications for children's development and language change*. PhD dissertation, University of Pennsylvania.
Ringe, Don & Joseph F. Eska (2013). *Historical linguistics: toward a twenty-first century reintegration*. Cambridge: Cambridge University Press.
Rogers, James, Jeffrey Heinz, Margaret Fero, Jeremy Hurst, Dakotah Lambert & Sean Wibel (2013). Cognitive and sub-regular complexity. In Glyn Morrill & Mark-Jan Nederhof (eds.) *Formal grammar*. Berlin: Springer, 90–108.
Rosenthall, Samuel (1989). *The phonology of nasal–obstruent sequences*. Master's thesis, McGill University.
Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport (1996). Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928.
Saffran, Jenny R., Elissa L. Newport, Richard N. Aslin, Rachel A. Tunick & Sandra Barrueco (1997). Incidental language learning: listening (and learning) out of the corner of your ear. *Psychological Science* **8**, 101–105.
Sanders, Robert Nathaniel (2003). *Opacity and sound change in the Polish lexicon*. PhD dissertation, University of California, Santa Cruz.
Santelmann, Lynn M. & Peter W. Jusczyk (1998). Sensitivity to discontinuous dependencies in language learners: evidence for limitations in processing space. *Cognition* **69**, 105–134.
Schuler, Kathryn D., Charles Yang & Elissa L. Newport (2016). Testing the Tolerance Principle: children form productive rules when it is more computationally efficient to do so. *Proceedings of the Annual Meeting of the Cognitive Science Society* **38**, 2321–2326.
Seidl, Amanda & Eugene Buckley (2005). On the learning of arbitrary phonological rules. *Language Learning and Development* **1**, 289–316.
Smith, Nelson V. (1973). *The acquisition of phonology: a case study*. Cambridge: Cambridge University Press.
Smolensky, Paul (1996). The initial state and 'richness of the base' in Optimality Theory. Technical Report JHU-CogSci-96-4, Department of Cognitive Science, Johns Hopkins University.
Smolensky, Paul & Géraldine Legendre (2006). *The harmonic mind: from neural computation to Optimality-Theoretic grammar, volume I: cognitive architecture*. Cambridge, MA: MIT Press.
Sutskever, Ilya, Oriol Vinyals & Quoc V. Le (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume **2**. Cambridge, MA, MIT Press, 3104–3112.
Tesar, Bruce & Paul Smolensky (1998). Learnability in Optimality Theory. *LI* **29**, 229–268.
van de Vijver, Ruben & Dinah Baer-Henney (2014). Developing biases. *Frontiers in Psychology* **5**, Article No. 634.
Weide, Robert (2014). The Carnegie Mellon University pronouncing dictionary, v. 0.7b. Available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.
White, James, René Kager, Tal Linzen, Giorgos Markopoulos, Alexander Martin, Andrew Nevins, Sharon Peperkamp, Krisztina Polgárdi, Nina Topintzi & Ruben van de Vijver (2018). Preference for locality is affected by the prefix/suffix asymmetry: evidence from artificial language learning. *NELS* **48**, 207–220.
White, Katherine S., Sharon Peperkamp, Cecilia Kirk & James L. Morgan (2008). Rapid acquisition of phonological alternations by infants. *Cognition* **107**, 238–265.
Wiese, Richard (1996). *The phonology of German*. Oxford: Clarendon.
Yang, Charles (2016). *The price of linguistic productivity: how children learn to break the rules of language*. Cambridge, MA: MIT Press.
Yang, Charles, Stephen Crain, Robert C. Berwick, Noam Chomsky & Johan J. Bolhuis (2017). The growth of language: Universal Grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews* **81**, 103–119.

https://doi.org/10.1017/S0952675724000083 Published online by Cambridge University Press