Learning phonological underlying representations: the role of abstractness*

Yang Wang
University of Utah

Bruce Hayes
UCLA

To appear, in copy-edited form, in *Linguistic Inquiry*

Abstract

We explore a novel approach to learning underlying representations (URs) which incorporates a number of current proposals in phonological theory and computational modeling. We seek to bring our results to bear on the long-standing issue of abstractness in phonology. Our strategy is to run the same learning model on a variety of languages while systematically varying the degree of abstractness permitted, following the abstractness hierarchy set forth by Kenstowicz and Kisseberth (1977). We find that when the criterion of abstractness is permissive, the resulting large set of candidate URs can lead the learning system to fail by getting stuck in a local maximum. We invoke research suggesting that abstract systems are often mislearned by children, and identify a level of Kenstowicz and Kisseberth's abstractness hierarchy that best predicts such restructurings.

Keywords: underlying representation, abstractness, learnability, alternation learning, Catalan, Tangale, Seediq

* We would like to thank two anonymous reviewers for *Linguistic Inquiry*, Adam Albright, Arto Anttila, Eric Baković, Jane Chandlee, Tim Hunter, Adam Jardine, Gaja Jarosz, Claire Moore-Cantwell, Jason Eisner, Charlie O'Hara, Paul Kiparsky, Adeline Tan, Colin Wilson, Tony Yates, members of the UCLA Phonology Seminar, and audiences at Cornell University, Seoul National University, Stony Brook University and AMP 2022 for helpful discussion. Jennifer Kuo both provided helpful advice and shared her data files for Seediq.

The Supplementary Materials mentioned at various places in the text may be obtained from https://osf.io/abktn/?view_only=2c34ea0f2dff4184abc831c16a895c44.

1. Stating the problem

We seek to contribute to a growing body of research attempting to solve the problem of morphophonemic learning, specifically the learning of underlying representations along with the grammar that maps them into surface forms. Characteristically, researchers have employed training sets of morphological paradigms, and developed systems that can recover the underlying representation (UR) for each morpheme, along with the phonological grammar whereby surface forms are derived.

This has proven to be an interesting challenge, and has been undertaken with a wide variety of approaches: error-driven learning with ranked constraints (e.g. Tesar et al. 2003; Apoussidou 2007; Merchant 2008; Tesar 2014; Nyman and Tesar 2019); Minimum Description Length with constraints (Rasin and Katzir 2016) and with rules (Rasin and Katzir 2018, Rasin and Katzir 2020, Rasin et al. 2021); formal learning algorithms based on subregular phonology (Hua and Jardine 2021); Bayesian Program Synthesis with ordered rules (Barke et al. 2019, Ellis et al. 2022), algorithmic learning guided by different evaluation criteria (Khalifa et al. 2023, Belth 2023), and maximum likelihood learning under different probabilistic frameworks (Jarosz 2006; Pater et al. 2012; Cotterell et al. 2015; Johnson et al. 2015; O'Hara 2017; Nelson 2019; Tan 2022).[2]

In this work we incorporate many ideas from this research tradition. We differ, however, in one of our research goals: we hope to bring computational learning theory to bear on a long-standing issue in theoretical phonology: the "abstractness controversy," which addresses the question of how far URs can diverge from surface representations (SR). This question has been debated on empirical grounds, with scholars attempting either to justify "deep" URs for some language, or contrariwise to argue that URs beyond a certain degree of abstractness are never entertained by human learners. For our review of the controversy, see §2.

In earlier research on learning, authors have assumed a particular position on the controversy, be it concrete or abstract, and let this assumption be an invariant property of the learning model. In contrast, in our own learning system, the degree of abstractness permitted is varied systematically. When we apply our learning system to language data, we keep the basic learning mechanism constant, while varying the principles that regulate the abstractness of URs. Our interest is in cases in which the same learning system succeeds under one approach to abstractness, but fails at another, thus potentially shedding light on the abstractness controversy.

Our research has benefited from the study of a classic work that put forth a taxonomy of UR-SR distance, Kenstowicz and Kisseberth (1977, ch. 1). This taxonomy consists of a hierarchical series of criteria governing UR-SR distance, which we will call the **KK**

[2] The field has grown so large that it is difficult to list all relevant work; for more complete literature survey, see Jarosz (2013, 2019), Tesar (2014), Cotterell et al (2015), and Rasin et al. (2021).

Hierarchy. We will use a subset of it as the basis for evaluating learnability implications of abstractness.[3]

We find that several levels of the KK Hierarchy are construable as algorithms for creation of UR candidate sets. Given a set of allomorphs for the set of morphemes under consideration, each successively higher KK level can generate a larger set of candidate URs. To preview an example from §6.3, the Seediq stem for 'hold' surfaces as either [pemux] or [pumex]. From these allomorphs, our implementation of KK's "level D" generates the following set of candidate URs: {/pemex/, /pemux/, /pumex/, /pumux/}. We show that given such candidate sets and appropriate constraints, it is feasible to learn a phonological system consisting of URs and constraint weightings that generates the Seediq paradigms. We express three other levels (KK-B″, KK-C, and KK-E) as UR candidate-creation algorithms.

Our work ultimately leads to a demonstration example (§6.6) in which a language ceases to be learnable when we move too high on the KK Hierarchy — the search space is so large that our system gets stuck in a local optimum. In this way, we show that, at least in principle, the abstractness problem can be addressed with computational implementations.

The rest of the article is organized as follows. We first discuss phonological abstractness (§2), the KK Hierarchy (§3), and the toy data set used to illustrate our system (§4). §5 gives the architecture of our learning system, and §6 applies it to a series of examples at varying levels of the KK Hierarchy, culminating in an example in which going too high blocks learning. Lastly (§7), we offer our conclusions and directions for future research.

2. The abstractness debate: where do things stand?

The research era of *SPE* (Chomsky and Halle 1968) was notable for the degree of abstractness it proposed to tolerate in underlying forms: *SPE* employed pre-Vowel Shift URs for English long vowels (/diviːn/ for [dɪˈvaɪn] *divine*); Foley (1965) used /eː, oː/ for the Spanish [e, o] that don't diphthongize under stress; Lightner (1965) used /ɪ, ʊ/ for the Russian vowels that alternate as [e/o ~ ∅]; and so on.

In reaction, the "abstractness debate" began: scholars questioned whether highly abstract analyses match what native speakers really learn; see e.g. Derwing (1973), Hooper (1976), and Tranel (1981). The work of Kiparsky (1968, 1973, 1982) was particularly influential; it outlined some possible limitations on abstractness backed by

[3] Kenstowicz and Kisseberth's original purpose was actually not the same as ours; they sought to demonstrate that there is essentially *no* limit to how much URs may diverge from SRs in a descriptively-adequate phonological analysis. However, at the end of their discussion (pp. 61-62), they noted that their conclusion was provisional, pending further research on whether abstract analyses are really internalized by learners. It is from this perspective that we return to their ideas.

empirical arguments from language change. However, not all scholars adopted concretist views, and defenses of abstractness continued (e.g. Gussmann 1980, Dresher 1981). A helpful review of the early debate is Sommerstein (1977); for a recent update see Baković et al. (2022).

Since the original controversy, new research has been carried out, including on some of the languages adduced in support of abstractness. Such work suggests to us that an approach that rejects extreme abstractness is likely to be on the right track. We follow up on two points made in the early debates.

No "resurrections" of abstract URs. An argument made by Kiparsky (1973:26-27) seems as valid today as when he made it: in historical change, abstract URs are never "resurrected" as actual pronunciations. Thus, there is no Slavic dialect in which Lightner's [ɪ], [ʊ] have been restored as surface segments; no variety of Hungarian where the [iː]'s that trigger exceptional back harmony have reverted to the [ɯː] proposed for them by Vago (1976) and others. A simple explanation for why language learners never come to pronounce highly abstract URs faithfully is that they do not entertain them during learning.

Abstractness-amenable phonological patterns are unstable. As preface we observe that, as numerous scholars have noted (see Hayes and White 2015 for a recent review), a phonological pattern amenable to an abstract analysis generally is the result of diachronic change. New phonological changes, applying "postlexically" as in Kiparsky (1982), apply to obscure an older phonological pattern, and the abstract analyst recapitulates this history as synchrony. The key question, however, is what happens when a new generation of language learners must make sense of the new pattern. As Kiparsky showed (1973:27-28), there is evidence that language learners who confront "abstractness-friendly" data patterns tend to remold their language, replacing the older data pattern with novel, concreteness-compatible patterns; a process called *restructuring* (Bynon, 1977). For instance, Kiparsky (1973:27-28) addresses a Mongolian example: the sound change ɯ > i created stems with surface [i] that take back harmony. In later change, such stems evolved to take uniform front harmony, rendering the /ɯ/ hypothesis vacuous.

Subsequent research has yielded similar findings. Hansson and Sprouse (1999) show that Yowlumne (Yawelmani) evolved in notably concretist directions during the decades between Newman (1944) and their own fieldwork. Polish jer alternations have evolved from a system recapitulated as synchrony by Gussmann (1980) and Rubach (1984) to a concretist system based on sonority sequencing (see Gorecka 1988, Czaykowska-Higgins 1988, Jarosz 2008, Rysling 2016; and for Russian: Gouskova 2012, Rysling 2016). Brame's (1972) abstract-/ʕ/ analysis of Maltese was thrown into doubt by later observations of Comrie (1986). Hoffman's (1973) analysis of Okpe with abstract /ɪ, ʊ/ was refuted by Archangeli and Pulleyblank's (1994) phonetic evidence that these sounds are actually realized in surface forms. For related Urhobo, Aziza (2008) sets forth a system with abstract /ɪ, ʊ, ə/, but observes (p. 17) that for younger speakers it is breaking down. Even a rather modest form of abstractness — "composite URs," — likewise triggers reanalysis; see discussion in §7.1.2.

Another change since the heyday of the abstractness debate has been a new interest in phonological learning, studied with computational models. This, too, impinges on abstractness: abstract representations are not necessarily objectionable on their own terms, but need justification, especially for how they might be acquired (Dresher 1981, Odden 2005:297, Baković 2009). Computational exploration of the conditions under which abstract URs are learnable can address this question more directly. Indeed, the research of O'Hara (2017), Rasin and Katzir (2018), Nyman (2021), and Belth (2023) offers instances where — provided certain very specific conditions are met — abstract URs can be learned.

Our own results, it will emerge, are more circumspect: under the system we propose, there are benefits to using somewhat concrete representations. The key idea is to run the same learning system under a variety of different conditions, each corresponding to a different degree of abstractness being tolerated. This variety is provided for us by the KK Hierarchy.

3. The KK Hierarchy in outline

The lowest level of Kenstowicz and Kisseberth's hierarchy, KK-A, will concern us here only briefly; it requires URs to contain "all and only the invariant phonetic properties of [a] morpheme." This criterion is so strict as to forbid ordinary phonemicization; for example the UR of English *pan* [pæ̃n] would have to be /pæ̃n/, including allophonic nasalization. We agree with Kenstowicz and Kisseberth (1977:9-11) that this is probably too restrictive. In what follows we assume that UR learning makes use of representations that have been phonemicized; our system as applied to English would "see" [pæn], not [pæ̃n]. This lets the system engage with a much smaller segment inventory without loss of essential information. For textbook background on phonemicization see Hayes (2008:20-28), and for current computational models that can phonemicize see Peperkamp et al. (2006), Calamaro and Jarosz (2015), Rasin et al. (2021), and Richter (2021).

Our main focus will be on four higher levels of the KK Hierarchy, all of which presuppose phonemicization.[4]

[4] KK propose four additional levels. At KK-A′ only invariant features are listed, with alternating features left unspecified. KK-B requires that the UR be based on the isolation form. KK-B′ says to pick the UR from the allomorph that occurs in the most contexts. We agree with KK's arguments that these three levels are too restrictive to be viable. KK-F says that at for each segment in a morpheme, at least one feature value from some surface form of that segment must appear in the UR. KK-F permits vast numbers of UR candidates and we have not attempted to test it.

KK-B″: The single-surface base hypothesis. In KK-B″, the language learner chooses a particular paradigm slot (e.g., for Yiddish verbs, the 1st singular) and employs its allomorphs as the URs. This highly restrictive hypothesis has since been pursued by Albright and others, who demonstrate its capacity to explain historical change in Yiddish (Albright 2010), Lakhota (Albright 2002), Latin (Albright 2005), Pengo (Dowd 2005), and Korean (Kang 2006, Albright 2008). The KK-B″ hypothesis is so restrictive one might wonder why a learning system would even be useful; we give a reason in §7.1.4.

KK-C: Choose among allomorphs. The UR of a morpheme is always identified with *some* allomorph in the morpheme's paradigm, but that allomorph need not come from the same paradigmatic slot for every morpheme. Thus, the UR candidate set is identical to the phonemicized allomorph set.

KK-D: Segmentally-composite URs. Every segment in the UR must be realized faithfully in *some* surface form, but not necessarily the same form, so URs can be "cobbled together" from multiple allomorphs. Thus, in *SPE* the UR /tɛlɛgɹæf/ is formed from *telegraph* [ˈt**ɛ**ləˌgɹ**æ**f] and *telegraphy* [təˈl**ɛ**gɹəf-i]. Because unstressed vowels are neutralized to schwa, no single allomorph informs us about all three underlying vowels.

KK-E: Featurally-composite URs. This level permits URs to contain segments never present in the surface allomorphs, so long as every *feature* within them occurs in some allomorph. Yates (2017) suggests Hittite phonology as an example of KK-E: he posits UR vowels that are [+long, −accent]; these always surface either as [−long, −accent] or as [+long, +accent], but never in their unchanged UR form.

We note that the hypotheses embodied in these levels are not about languages, but about the language faculty: at most one of them can be true; and all must be tested against multiple languages for their impact on learnability.

4. A toy example for illustration: Pseudo-German

We illustrate our UR learning system using a simple data example, modeled on Pater et al. (2012), to be called Pseudo-German. We give the data in the form that the learner sees: phonemic transcription, glosses indicating what morphemes are present, but no morpheme boundaries per se. Stems are identified as such in the glosses. The glosses are given in correct linear order, but the model is not informed of this order.

(1) *Pseudo-German input data*

a.	[be**t**]	cat$_{stem}$	[be**d**a]	cat$_{stem}$ plur.
b.	[mot]	dog$_{stem}$	[mota]	dog$_{stem}$ plur.
c.	[lo**p**]	turtle$_{stem}$	[lo**b**a]	turtle$_{stem}$ plur.
d.	[pap]	soup$_{stem}$	[papa]	soup$_{stem}$ plur.
e.	[mi**k**]	plane$_{stem}$	[mi**g**a]	plane$_{stem}$ plur.
f.	[bek]	beer$_{stem}$	[beka]	beer$_{stem}$ plur.

g.	[es]	$\text{wine}_{\text{stem}}$	[esa]	$\text{wine}_{\text{stem}}$ plur.
h.	[nur]	$\text{light}_{\text{stem}}$	[nura]	$\text{light}_{\text{stem}}$ plur.
i.	[to]	toe_{stem}	[toa]	toe_{stem} plur.

In Pseudo-German, the alternating stems have final voiced obstruents underlyingly (/bed/, /lob/, /mig/), and their isolation forms are derived with a phonology that devoices obstruents word-finally. The remaining stems do not alternate and may be assigned URs identical to their phonetic form. Forms like [mota] 'dog-plur.' show that the alternations seen in 'cat' etc. cannot be attributed to a process of intervocalic voicing.

5. Description of the learning system

In earlier research on computational morphophonemic learning, scholars have prioritized a variety of different goals. Here are three goals to which we aspire here.

Scale of analysis. In a subset of the literature, scholars have tested out their analyses on very small, schematic data sets. While this approach has undeniable advantages for understanding how the model functions, an increasing trend has been to test models against more substantial data, at least on the scale of an ordinary problem set, with multiple phenomena addressed — see e.g. Cotterell et al. (2015), Barke et al. (2019), Ellis et al. (2022), and Belth (2023). Such work increases our confidence that the model can ultimately be scaled up to match the experience of human children.

Use of a constraint-based framework. Many models have employed a grammatical system that uses constraints and a GEN-cum-EVAL architecture. This type of model originated in classical Optimality Theory (OT; Prince and Smolensky 1993) and has supported much research on UR learning (Tesar and Smolensky 2000; Jarosz 2006; Apoussidou 2007; Tesar 2014; Cotterell et al. 2015; Rasin and Katzir 2016). We opine that constraint-based frameworks such as OT are likely to have more explanatory value than models based on rules or other schemata. The constraints of OT relate the content of language-specific grammars to broader principles grounded in typology and phonetics. Further, OT has proven well suited to the incorporation of **biases** (e.g., simplicity, phonetic naturalness, paradigm uniformity) into phonological learning; see e.g. Tesar and Smolensky (2000), Wilson (2006), Zuraw (2007, 2013), Becker et al. (2012), White (2017), and Kuo (2024).

Use of probability. Classical OT provides no basis for the analysis of gradient phenomena, such as free variation (Labov 1969 et seq.), lexical frequency matching (Zuraw 2000 et seq.), and the soft-UG learning biases just noted. For this reason, starting with Jarosz (2006), systems for UR learning have often employed probabilistic versions of OT. Here, we adopt Maximum Entropy grammars ("MaxEnt"; Goldwater and Johnson 2003). MaxEnt inherits its key elements from OT (GEN-cum-EVAL architecture, constraint system), but instead of ranking the constraints it assigns every constraint a numerical weight. On the basis of the weights and the pattern of constraint violations, a probability is computed for each candidate (see §5.5 below). MaxEnt is computationally tractable and is used in several current systems for UR learning (Pater et al. 2012, Johnson et al. 2015, O'Hara 2017, Nelson 2019, and Tan 2022).

Our other use of probability is in the choice of UR: competing UR candidates receive probabilities, with the correct one normally achieving probability very close to 1 by the end of learning.

We factor the problem of learning URs into stages, as follows. The learner first assigns the segments of each input word form to their morphemes, thereby defining a set of allomorphs for each morpheme. Based on these allomorph sets, it detects the phonological alternations and constructs a set of candidate URs, according to whatever KK level has been chosen. The inputs to the system also include a set of hand-supplied constraints, to which the system assigns weights, forming a MaxEnt phonological grammar. The final task is to simultaneously determine the URs and constraint weights. Success is defined by a combination of these choices that accurately matches the frequencies of the training set.

5.1 Morpheme segmentation

The task of dividing a word (given in surface form) into its component morphemes has a substantial literature; for overview see Hammarström and Borin (2010). In the present context, we need a model that fits the following criteria. (1) It employs labeled training data, as in (1), with the morphosyntactic features of the forms specified. (2) It is not defeated by alternation; e.g. for Pseudo-German [beda] 'cat-pl.', the system can detect the stem [bed] even though it is not identical to the singular form [bet]. (3) The system should parse pre-phonologically, rather than trying to solve the parsing problem jointly with phonology, as with Jarosz (2006), Cotterell et al. (2015), Nelson (2019), and Rasin et al. (2021).[5] The capacity to parse pre-phonologically is needed in any event to detect irregular allomorphy (e.g. Korean nominative *-ka*/*-i*), where alternation occurs but is not attributable to general phonology.

Oddly, we have found no model that performs the learning task just described. We think that undertaking this task would be sensible, since there are several models of UR discovery (e.g., Tesar and Smolensky 2000, O'Hara 2017, Nyman and Tesar 2019, Belth 2023) that simply presuppose that a parse into morphemes has been carried out.

The model we have assembled is founded on the principle of **paradigm uniformity** (Kiparsky 1968, 1971; Steriade 2000; Wilson 2006; Zuraw 2007): the allomorphs of a morpheme within its paradigm tend to be phonetically similar. Even when we do not know in advance what the phonology will be, there is good reason to think that the alternations it imposes on morphemes will be relatively modest, so that bad parses reveal themselves by exhibiting excessive alternation. As an example, consider an erroneous

[5] For evidence that infants can perform morphological segmentation before they learn the system of phonological alternations, see Marquis and Shi (2012), Ladányi et al., (2020), and Sundara et al. (2021).

parse for Pseudo-German [beda] ‘cat-plural’ in which ‘cat’ is taken to be [be] and ‘plural’ [-da]. The error of this parse is detectible because it creates unnecessary dissimilarities among allomorphs: [be] for ‘cat’ is strongly dissimilar to the isolation allomorph [bet], and [-da] for ‘plural’ is dissimilar to the plural allomorphs that will arise in the paradigms of other stems. In contrast, where the data are correctly parsed ([bet-a]), dissimilarity is minimized, just the one-feature voicing difference between [d] and [t] in ‘cat’.

Our system searches over all the possible morpheme parses for the words in the training set, seeking the one that minimizes dissimilarity among allomorphs. We find that — at least for the examples considered in this article — the result of this procedure matches the linguist-preferred parse, and renders possible the later discovery of appropriate URs and phonology. We give a full description of the morpheme parser in Supplementary Materials 1.1.

5.2 Allomorph dictionary and morpheme ordering

The next step, following e.g. Tesar et al. (2003), Merchant (2008), and Tesar (2014), is to collect an allomorph set for every morpheme. These allomorphs are used to construct candidate URs, in ways that depend on the choice of KK level. To find the allomorph set for any given morpheme, it suffices to examine each word containing the morpheme and extract the strings of segments assigned under parsing to that morpheme. Following this procedure for all morphemes, we obtain an **allomorph dictionary**, illustrated in (2).

(2) *A partial allomorph dictionary for Pseudo-German*

‘cat’	{[bet], [bed]}
‘dog’	{[mot]}
‘plur.’	{[a]}

From such a list, it is usually straightforward to determine the principles of morpheme ordering.[6]

5.3 Detection of alternating segments

To learn the phonology, it is necessary to discover what segments alternate. Here we make use of optimized string alignment. For instance, an intuitively optimal alignment of the segments of Pseudo-German [bet] and [bed] is illustrated in (3a), and the intuitively optimal alignment in a case of epenthesis or syncope (segment paired with null) is illustrated in (3b).

[6] Since we assume uninterrupted morphemes, the following suffices: we find the set S_1 of morphemes that are never preceded by any other morpheme, then the set S_2 whose members are only ever preceded by members of S_1, and so on until every morpheme belongs to a set.

(3) *Two representative optimal allomorph alignments*

a. Pseudo-German 'dog'

b	e	d
b	e	t

b. Epenthesis/Syncope

a	p	t	ə
a	p	t	∅

In contrast, a non-optimal alignment would be [b]-∅, [e]-[b], [d]-[e], [∅]-[t] for (3a), with each segment of [bet] shifted one position to the right. The key idea is that where the optimal alignment pairs non-identical elements ((3a): [t]-[d], (3b): [ə] - ∅), these elements will generally be in phonological alternation.

To find the optimal alignment, we adopt a standard algorithm (Kruskal 1983), previously used for phonology by Bailey and Hahn (2001), Albright and Hayes (2003), and Moore-Cantwell and Staubs (2014). We have found in the cases we have examined that this method of detecting alternations is quite reliable. As applied to the Pseudo-German dataset of (1), the method yields the alternation set in (4):

(4) *Alternation set discovered for Pseudo-German*

[t] ~ [d]	(from [bet] ~ [bed], (3a))
[p] ~ [b]	(from [lop] ~ [lob])
[k] ~ [g]	(from [mik] ~ [mig])

When the learner first detects an alternation, it is agnostic regarding the appropriate phonological analysis. For Pseudo-German, the alternations in (4) must ultimately be attributed to final devoicing, not intervocalic voicing; but at this stage, all the learner knows is that [t] and [d] alternate.

5.4 Formation of candidate UR set

The learner next uses the discovered allomorphs to implement a chosen level of the KK Hierarchy. In KK-C, for instance, the UR candidates are simply the allomorphs themselves,[7] yielding (5) for Pseudo-German.

(5) *Sample UR candidates for Pseudo-German, KK level C*

'cat'	{/bet/, /bed/}
'dog'	{/mot/}
'plur.'	{/a/}

For higher levels, the choice of UR candidates is more complex, and we defer presentation until the particular examples of §6 that illustrate these levels. In most cases

[7] For discussion of KK-B″ see §7.1.4.

the KK levels are hierarchical, in that each higher level of the KK Hierarchy defines a candidate set that is equal to or a superset of the previous level.[8]

5.4.1 Probability distributions over URs

The UR candidates are in competition, and by the end of learning, one of them will usually dominate over the others.[9] Following earlier work (Jarosz 2006, Cotterell et al. 2015, O'Hara 2017) we represent this with a probability distribution over all UR candidates affiliated with a particular morpheme, summing to one.[10] This distribution is altered over the course of learning. Table (6) gives the initial and final probabilities assigned to representative candidate URs for our Pseudo-German learning simulation.

(6) *Initial and final (learned) probability distributions over URs for Pseudo-German*

Morpheme	*UR candidate*	*Initial probability*	*Probability after learning*
'cat'	/bet/	0.5	0
	/bed/	0.5	1
'dog'	/mot/	1	1
'plur.'	/a/	1	1

As can be seen, we set the initial probabilities assigned to rival candidate URs as equal.

Once created, the UR candidates for relevant morphemes are concatenated to create candidate URs for whole words. The morpheme ordering employed is the one obtained in §5.2. The probability of a candidate word UR is the product of the probabilities of the UR candidates of the morphemes that comprise it. Thus, for the Pseudo-German word [beda], there are two UR candidates, formed by concatenting the two UR candidates for the stem, /bed/ and /bet/, with the single candidate UR for the suffix, /-a/. These each bear the probability 0.5 × 1 = 0.5.

5.5 Phonological framework

We express the phonology using a MaxEnt OT grammar (Goldwater and Johnson 2003), in which the intuitive "strength" of constraints is represented with numerical weights rather than with ranking. For each language, appropriate constraints (mostly taken from

[8] The inclusion relations for all levels mentioned in the article are: B″ ⊆ C ⊆ D ⊆ E, D⊆ Z, Z⊆ EZ, E ⊆ EZ.

[9] When multiple URs each can derive the correct outcome, our system generally assigns each of them a non-zero probability.

[10] This is not the only way to represent UR choice probabilistically; an alternative is UR constraints (Apoussidou 2007, Eisenstat 2009, Pater et al. 2012, Johnson et al. 2015, Nazarov and Pater 2017, Nelson 2019), which prefer a particular UR and may be ranked or weighted with respect to the phonological constraints. As Cotterell et al. (2015) note, the probability distribution approach tends to favor a single UR for each morpheme and thus allocates more of the descriptive burden to the phonology.

the OT literature) are given to the learner in advance; for instance, in Pseudo-German, we hand-fed the model the constraints used by Pater et al. (2012): *FINAL VOICED OBSTRUENT (abbreviation *d]), *INTERVOCALIC VOICELESS OBSTRUENT (*VTV), and IDENT(voice). We discuss in §7.2.4 how the system might instead be able to learn its own constraints.

Below, we give a simple MaxEnt tableau, with just two candidates. It employs the constraint weights eventually learned by our system and assigns a suitably high probability (very near 1) to the correct candidate for /bed/, namely [bet]:

(7) *A MaxEnt tableau for* /bed/ → [bet]

/bed/	*d] w = 19.05	ID(voice) w = 10.09	$\mathcal{H}$	$e\mathcal{H}$	Z	p
a. [bed]	*		19.05	5.33×10^{-9}	4.15×10^{-5}	0.0001
b. ☞ [bet]		*	10.09	4.15×10^{-5}	4.15×10^{-5}	0.9998

Here, Harmony ($\mathcal{H}$) is the weighted sum of all constraint violations; $e\mathcal{H}$ is $exp(-\mathcal{H})$; Z sums $e\mathcal{H}$ over all candidates, and probability (p) for each candidate is its share in Z. The procedure for MaxEnt probability appears in Supplementary Materials 1.2.

We adopt here a particular stance on how the learning of alternations is related to phonotactics: following Hayes and Wilson (2008:§9.3) we assume that the phonotactic system is a separate component of the phonology; the phonotactic grammar is *not* the same as the alternation grammar, as classical OT assumes. Rather, the connection is based on learning: phonotactic learning takes place early (Jusczyk et al. 1993, et seq.), and the constraints that it yields serve as a source of hypotheses when, later on, the child turns to the learning of alternations (see Chong 2019, 2021).

We believe there may be advantages for learnability in thus severing the tight link made by classical OT between phonotactics and alternations. Notably, the GEN function for alternation can be simplified (§5.6), resulting in a smaller search space. Moreover, the many cases where phonotactics and alternation are actually misaligned become unproblematic. Paster (2013) divides these into two types: derived-environment processes (Kiparsky 1973) and stem-bounded phonotactics; see Paster for examples of the latter.

5.6 GEN for surface candidates

Our GEN function, inspired by Eisenstat (2009), is specifically suited to the learning of alternations; we call it **alternation-substitution**. It is based on the list of segmental alternations, as obtained in §5.3. For each UR candidate, GEN is obtained by applying every change on the alternation list wherever it is applicable, in all possible combinations. For instance, since [t] ~ [d] is in the alternation list for Pseudo-German, then for the UR /bed/, we must include the surface candidate [bet]. Similarly, for underlying /mot-a/ 'dog-pl.' we must include the losing candidate *[mod-a]. This incorrect candidate will prove informative, since it helps rule out an erroneous analysis with intervocalic voicing.

The process of alternation-substitution is illustrated for Pseudo-German below; note that incorrect candidate URs, such as /bet/, must also be assigned surface candidates.

(8) *GEN using alternation-substitution in Pseudo-German*

UR	*Applicable alternations (from (4))*	*SR candidates*
/mot-a/	[t] ~ [d]	[mot-a], [mod-a]
/bed/	[t] ~ [d], [p] ~ [b]	[bed], [bet], [ped], [pet]
/bet/ (incorrect)	[t] ~ [d], [p] ~ [b]	[bed], [bet], [ped], [pet]

In principle, free combination might lead to rather large GEN functions, since candidate count is the product of the number of possible alternating partners for every segment that a word contains. In the cases studied here, candidate counts are sometimes large (in the hundreds), but not so large as to make computation difficult.

Epenthesis requires special treatment, in order to keep the candidate set finite. For example, in English [ə] is used to resolve sibilant clusters arising in plurals, etc.: /fɪʃ-z/ → [fɪʃəz] 'fishes'. If we freely extended substitution of [ə] for zero, the candidate set would become infinite ([fɪʃəəz], [əəfɪʃəəəz], etc.). Our treatment works as follows. The [ə] ~ Ø alternation is first discovered through allomorph alignment (§5.3), as shown in (9):

(9) *Detecting epenthesis in the English plural suffix*

ə	z	as in *fishes* [fɪʃ<u>əz</u>]
Ø	z	as in *dogs* [dɔg<u>z</u>]

The system observes that there is a [ə] ~ Ø alternation in the local context / + ___ z (extracted from the full representation fɪʃ **+ ___ z**; + is a morpheme boundary). It then generalizes on the basis of local contexts (neighboring segments and boundaries). If the local context / + ___ z arises in an existing candidate, GEN is authorized to insert a [ə] to to create an additional candidate; thus, for /fɪʃ + z/, GEN would provide the correct candidate [fɪʃ + əz], even if the '*s* suffix had never been encountered before. This approach provides a sufficiently rich set of candidates to treat epenthesis without yielding an infinite set. This system must be considered provisional,[11] but it suffices for the examples considered.

[11] What is missing is the ability to generalize across feature-defined categories. Thus, if a language splits up vowel clusters with [ʔ], we expect it to do the same for novel vowels appearing in loanwords, which could not happen if GEN is guided by purely segmental contexts.

Our choice of an alternation-substitution GEN is dependent on our earlier choice (above) of a grammatical architecture in which phonotactics is treated separately from alternation. The reason is that in any language, there are phonotactic principles that are not enforced by alternations. Thus, Mandarin has no stop+liquid clusters, but the attested alternations of the language (3rd tone sandhi, etc.) have no bearing on this fact. Classical OT, with its Rich Base theory of phonotactics (Prince and Smolensky 1993:192, 209), requires a much larger GEN, with candidates to repair any input.

Our GEN system generally creates candidate sets smaller than those of other systems (see, e.g. Eisner 1997, Albro 2005, and Riggle 2004); these systems often can represent an infinite candidate set. However, Eisenstat's approach appears to be adequate to our purpose, and it embodies an intriguing empirical hypothesis (language learners are guided by the alternation set) that we feel is worth exploring further.

5.7 Finding the right combination of UR probabilities and constraint weights

The parameter sets whose values must be calculated are $\boldsymbol{\theta}$, the probability distributions over UR candidates, and $\boldsymbol{W}$, the weights of the phonological grammar. Simultaneous learning of $\boldsymbol{\theta}$ and $\boldsymbol{W}$ is the most challenging part of the whole model, for a reason pointed out in the literature (Tesar and Smolensky 2000; Jarosz 2019): the URs constitute a form of *hidden structure*. The choice of URs and the learnt phonology are mutually dependent, in that the choice of URs cannot be firmly established in the absence of a known phonological grammar, yet the weights of the phonological grammar themselves depend on the choice of URs.

The approach that we adopt is as follows. We employ an *objective function*, log likelihood, designed to reach its maximal value for the most accurate analysis. To find the values of $\boldsymbol{\theta}$ and $\boldsymbol{W}$ that maximize log likelihood, we employ *Expectation-Maximization* (EM, Dempster et al. 1977), an algorithm widely adopted (Jarosz 2006 et seq.) in phonological learning as a way of dealing with the hidden-structure problem. In EM, we first calculate the best-fit values for $\boldsymbol{\theta}$ employing a provisional estimate for $\boldsymbol{W}$. These $\boldsymbol{\theta}$ values in turn enable an improved estimate for the values of $\boldsymbol{W}$. The process continues, back and forth, until log likelihood ceases to improve by more than a small threshold amount.

Our implementation of EM, with the specific formulae used, is described in Supplementary Materials 1.2. For an illustration of how EM is applied in a specific case, see the spreadsheet for Pseudo-German in Supplementary Materials 2.

5.8 Results for Pseudo-German

The course of learning for Pseudo-German is given in (10).

(10) a. *Weights* (***W***)

Constraint	*Initial weight*	*Final learned weight*
*FINAL VOICED OBSTRUENT	1	19.05
*INTERVOCALIC VOICELESS OBSTRUENT	1	0.01
IDENT(voice)	1	10.09

b. *Sample UR probabilities* (***θ***)

	Initial probability	*Final Probability*
θ(‘cat’, /bed/)	0.5	>0.999
θ(‘cat’, /bet/)	0.5	0
θ(‘dog’, /mot/)	1	1

c. *Sample final probabilities for surface candidates*

Form		*SR candidate*	*Final probability*
‘cat’	☞	[bet]	> 0.999
		[bed]	~ 0
‘cat pl.’	☞	[bed-a]	> 0.999
		[bet-a]	~ 0

As can be seen, the learned values for ***θ*** and ***W*** suffice to select the linguist-expected URs and constraint weights with acceptable accuracy.

5.9 On local and global optima

EM resolves a complex optimization task into two simpler tasks, both of which search convex spaces and are thus guaranteed to find their own global optima. Specifically, at any stage, the constraint weights ***W*** are provably the global optimum for deriving the surface forms from the UR distributions assumed at that stage (Della Pietra et al. 1997); and the UR probabilities ***θ*** are likewise the optimum for deriving the surface forms given a set of constraint weights (McLachlan and Peel, 2000:47-50).

However, EM as a whole comes with no such guarantees: it provably arrives at *some* optimum in its search space, but this is sometimes a local optimum, not the global optimum. When the system we describe gets stuck in a local optimum, it fails to learn the correct grammar.

The tendency of EM models to get stuck in local optima has been noted as a disadvantage of this method (Cotterell et al. 2015). However, in a research context, this tendency might be an advantage (Nazarov and Pater 2017), since it can be made to bear on the choice of the empirically correct KK level. We will develop this line of inquiry in §6.6 and §7.1.

6. Case studies

We turn now to the application of our model with case studies. Our research design was to try to learn the same set of languages under various KK levels, setting all other model parameters constant. The learning simulations are arranged in an order based on how high on the KK hierarchy one must ascend in order to learn the URs. For the settings we employed in running EM, see Supplementary Materials 1.2. We mostly rely on standard constraints taken from McCarthy and Prince (1995). For all of our examples, we offer documentation of how the software implementing our algorithms did its work; see Supplementary Materials 3.

6.1 Catalan phonology

Catalan (Romance, Catalonia and neighboring areas) is an ideal language for exploring UR learning at problem-set scale. Catalan phonology includes lexical variation and opaque process interaction, offering additional challenges to learning systems. We have been able to rely on an extensive body of descriptive and analytical work, including Mascaró (1976), Wheeler (2005), and Bonet and Lloret (2018).

6.1.1 Data and basic phonological analysis

Our data roughly match what was used by Cotterell et al. (2015), who drew on Kenstowicz and Kisseberth's (1979) textbook. In order to include lexical variation, we added additional paradigms from the sources cited above.[12] Our full set consists of 33 four-member paradigms in which stems are inflected for both gender (masculine -∅ and feminine [-ə][13]) and number (singular and plural), as in (11). We include morpheme boundaries for clarity, though they are not present in the data presented to our system.

(11) *Representative paradigms from the Catalan training data*

	UR	*m.sg.*	*m.pl.*	*f.sg.*	*f.pl.*	
a.	/kru/	kru	kru-s	kru-ə	kru-ə-s	'raw'
b.	/ultim/	ultim	ultim-s	ultim-ə	ultim-ə-s	'last'
c.	/petit/	petit	petit-s	petit-ə	petit-ə-s	'small'
d.	/sek/	sek	sek-s	sek-ə	sek-ə-s	'dry'
e.	/sant/	san	san-s	sant-ə	sant-ə-s	'holy'
f.	/fort/	for	for-s	fort-ə	fort-ə-s	'strong'

[12] In a few cases the references had modest gaps, e.g. including the feminine singular but not the feminine plural. We added forms to fill these gaps, checking them first against online pedagogical sources for Catalan.

[13] On orthographic and historical grounds, the feminine singular might be treated as /-a/ and the feminine plural as /-es/, with [ə] derived from these URs by a well-motivated process of Vowel Reduction. As a test of our morpheme parser, we opted for an analysis in which the feminine vowel is uniform and the plural is [-s] added to the singular, as in the masculine.

g.	/bon/	bo	bon-s	bon-ə	bon-ə-s	‘good’
h.	/klar/	kla	kla-s	klar-ə	klar-ə-s	‘plain’
i.	/kazad/	kazat	kazat-s	kazad-ə	kazad-ə-s	‘married’
j.	/seg/	sek	sek-s	seg-ə	seg-ə-s	‘blind’
k.	/griz/	gris	griz-us	griz-ə	griz-ə-s	‘grey’
l.	/bɔʒ/	bɔtʃ	bɔʒ-us	bɔʒ-ə	bɔʒ-ə-s	‘crazy’
m.	/gros/	gros	gros-us	gros-ə	gros-ə-s	‘big’
n.	/despatʃ/	despatʃ	despatʃ-us	despatʃ-ə	despatʃ-ə-s	‘office’

Generally, stems appear unaltered in the feminine forms,[14] whereas in the masculine forms various types of phonology attack the exposed coda consonants.

Final cluster simplification. This can be seen in (11e) and (11f), which illustrate the simplification of /nt/ to [n] and /rt/ to [r] when a word boundary or consonant follows. More generally, the stem-final clusters that simplify are, roughly, the homorganic ones; for more details and closer analysis see Wheeler (2005:220-235). In our own analysis, we assume that a constraint to the effect of *FINAL HOMORGANIC CLUSTER has a much higher weight than MAX, so that simplification is essentially obligatory. We avoid bad solutions like /sant/ → *[sat] by assigning a high weight to I-CONTIGUITY-STEM (McCarthy and Prince 1995), which bans skipping.

Deletion of singleton /n/ and /r/. Two singleton consonants, /n/ in word-final position and /r/ in coda position, are also targeted for deletion; see (11g) and (11h). As Wheeler (2005:327-338) emphasizes, the loss of singleton consonants is not a regular, across-the-board process, but involves many exceptions. We include in our training set both deleting and non-deleting stems for both /n/ and /r/, selecting (as an approximation) a ratio of three deleting stems for each non-deleting stem.[15]

(12) *Forms that drop, or do not drop, final /n/ and /r/*

a. *Lexically-specific* /n/ *Deletion*

m.sg.	*m.pl.*	*f.sg.*	*f.pl.*	*Gloss*
bo	bon-s	bon-ə	bon-ə-s	‘good’
ple	plen-s	plen-ə	plen-ə-s	‘full’
prəgo**n**	prəgon-s	prəgon-ə	prəgon-ə-s	‘proclamation’

[14] Feminines do involve Spirantization of the stem-final consonant, as in [kazað-ə-s] ‘married f.sg.’, stem UR /kazad/. As proposed in (§3), our inputs to learning consist-of phonemicized data, so that at the level of analysis discussed here the spirant allophones of Catalan ([β,ð,ɣ], Wheeler 2005:310-327) are represented with their phonemic values /b,d,g/.

[15] The lexical survey for Liang et al. (2024) suggests somewhat higher rates for /n/ deletion and /r/ deletion; about 90% and 93%, respectively. We use 75% for ease of diagnosis (75% is not close to 1), though we find that our model can also match higher frequencies.

b. *Lexically-specific* /r/ *Deletion*

m.sg.	*m.pl.*	*f.sg.*	*f.pl.*	*Gloss*
du	du-s	dur-ə	dur-ə-s	‘hard’
kla	kla-s	klar-ə	klar-ə-s	‘plain’
pu**r**	pu**r**-s	pur-ə	pur-ə-s	‘pure’

We agree with Wheeler (p. 330) that unpredictable forms in cases like this must be lexically listed in some way. Of particular interest to us is how native speakers would respond in circumstances requiring them to use their grammar productively, as in encounters with a novel stem, or a wug test. Wug-testing in other languages suggests that, all else being equal, adult speakers are likely to *frequency-match* such patterns (see e.g. Zuraw 2000, Ernestus and Baayen 2003); [16] and indeed Liang et al. (2024) demonstrate frequency matching for Catalan /n/ and /r/ Deletion across a variety of environments. Here we employ an idealized version of frequency matching, eschewing detailed environments and assuming a 75% deletion rate across the board.

/n/ Deletion and /r/ Deletion are not just exceptionful, but also opaque: as expressed in rule-based phonology, they are *counterfed* by Cluster Simplification. The grammar must guarantee this counterfeeding effect, since alternations like *[sant-ə] ~ [sa] are not possible. For this purpose, we adopt a highly weighted Faithfulness constraint banning alternations of the form *CC ~ ∅, following Kirchner (1996).

Epenthesis. The data in (11k)-(11n) illustrate a process of epenthesis, which splits up sibilant clusters created when the plural ending is attached to a sibilant-final stem, as in /griz-s/ → [grizus] ‘grey-m.pl.’ To treat this, we assume a constraint banning adjacent sibilants (*SIBILANT CLASH), with sufficient weight to overcome the Faithfulness constraint DEP. In real Catalan, the process is arguably not epenthesis but morphological in character (Wheeler 2005:263-264), but we assume epenthesis here in order to test our system’s ability to deal with such phenomena.

Devoicing in codas. Catalan has a devoicing process similar to Pseudo-German, illustrated in (11j)-(11m); for example /griz/ → [gris]. Cases like (11k) /seg-s/ → [sek-s] suggest that devoicing applies to all coda obstruents, not just word-final ones, so we adopt *CODA VOICED OBSTRUENT, which must outweigh IDENT(voice). We also included the useless constraint *VTV, to check that the system does not wrongly learn a system of intervocalic voicing.

An interesting related issue is that final /ʒ/ surfaces in final position not as the expected [ʃ] but as [tʃ], as in /bɔʒ/ → [bo**tʃ**], (11l). This is a case of “saltation” (Lubowicz

[16] More precisely, the observed output normally emerges as a compromise between frequency-matching and various UG biases; see e.g. Hayes et al. (2009) and Becker et al. (2012). Frequency-matching is also observed, but not as consistently, in children and in artificial grammar learning studies; see Hudson Kam and Newport (2009) and for a recent overview Schumacher and Pierrehumbert (2021).

2002, Ito and Mester 2003, Hayes and White 2015), in that the /ʒ/ ([+voice, +continuant]) leaps across phonetically intermediate [ʃ] ([−voice, +continuant]) in arriving at [tʃ] ([−voice, −continuant]). The phenomenon is discussed in detail (and analyzed using abstract phonology) by Bonet and Lloret (2018). Here, we follow the analytic approach of Hayes and White (2015), banning the expected-but-unwanted "short" journey using a *MAP constraint (Zuraw 2007, 2013), which bans a particular segmental correspondence. Here, *MAP(ʒ ~ ʃ), weighted higher than *MAP(ʒ ~ tʃ), forces the outcome [tʃ] in [bɔtʒ]. In Zuraw's theory, the weighting is an unnatural one, in that the shorter phonetic path is penalized more than the longer one. Below we discuss evidence that this unnatural pattern indeed tends to lead to repair by Catalan language learners. In the analysis, we included several *MAP constraints definable over the segments [ʃ, ʒ, tʃ, dʒ].

In sum, our phonological analysis is based on the constraints in (13), which are listed with the weights that were assigned to them by our learning system.

(13) *Constraints used for Catalan with their model-computed weights (KK-C)*

Name	*Function*	*Weight in model*
*FINAL HOMORGANIC CLUSTER	Triggers simplification of final clusters	39.89
MAX	Militates against deletion	15.85
*FINAL [n]	Triggers lexically variable /n/ deletion	16.95
*CODA [r]	Triggers lexically variable /r/ deletion	16.95
I-CONTIG-STEM	Forces deletion to be morpheme-peripheral	25.61
*CC ~ Ø	Forces counterfeeding relation between cluster simplification and singleton deletion	10.83
*SIBILANT CLASH	Triggers epenthesis in sibilant clusters	47.26
DEP	Militates against epenthesis	39.18
*CODA VOICED OBSTRUENT	Triggers Final Devoicing	20.80
*VTV	Useless, intended to challenge the system (don't discover Intervocalic Voicing)	0.00
IDENT(voice)	Faithfulness	10.13
IDENT(continuant)	Faithfulness	0.97
*MAP(ʒ ~ ʃ)	Forces saltation of /ʒ/ to [tʃ]	10.18
*MAP(ʒ ~ tʃ)	Must be weighted low (unnatural by P-map)	0.00
*MAP(ʒ ~ dʒ) *MAP(tʃ ~ dʒ) *MAP(tʃ ~ ʃ)	Other *MAP constraints defined on nonanterior sibilants	0.38 0.86 9.01

6.1.2 Results of our learning procedures for Catalan

We report the results obtained using KK-C (all and only allomorphs are UR candidates), but in this case the choice of level does not matter, since all four KK levels of §3 suffice to find the right analysis.[17]

Here is the sequence of events that yielded the correct outcome. During the initial stage of morpheme parsing (§5.1), the feminine suffix was correctly identified as [-ə] and the plural as [-s/-us]; other segments were correctly assigned to stems. An example of an output parse is [bon-ə-s], 'good-feminine-plural'. Following the compilation of allomorphs (§5.2), the correct order of morphemes was detected, namely stem, gender, number. The alignment procedure of §5.3 correctly located the segmental alternations, which are { n ~ ∅, r ~ ∅, d ~ ∅, t ~ ∅, k ~ ∅, p ~ ∅, u ~ ∅, b ~ p, d ~ t, g ~ k, z ~ s, dʒ ~ tʃ, ʒ ~ tʃ }. The candidate URs created under KK-C (and all other levels) included e.g. {/bon/, /bo/} for 'good'; there were always one or two candidate URs per stem. The learner used alternation-substitution (§5.6) to create surface GEN, yielding a total of 14,130 UR-SR pairs. The hand-provided constraints were as in (13). The constraints, candidate URs, and SR candidates for each UR were input to EM-based learning (§5.7), which estimated UR probabilities and constraint weights. The system met the convergence criterion after 25 iterations.

The learning run yielded the following results. (a) Underlying representations: For all 33 stems, the UR chosen matched the feminine stem allomorph, corresponding to hand analysis. In every case, the probability assigned to this UR was at least 0.999. (b) Constraint weights as in (13). (c) Probability assigned to surface candidates: for invariant forms, this was always at least 0.99 for the correct candidate. For the cases of 3-to-1 lexical variation, the majority candidate was assigned a probability between 0.748 and 0.749. In other words, the learned grammar achieved a close match to the patterns present in the training set.

The learner correctly attributed all variation to the grammar, rather than to multiple URs. For example, while 'good' surfaces as both [bo] and [bon], the learner assigned essentially 100% probability to /bon/, with [bo] derived by the phonology. This is sensible, since allocating any probability at all to /bo/ would derive incorrect results in the feminine forms.

We conclude that our learning system, given labeled data and a constraint inventory, was able to solve the Catalan problem at every level of description.

---

[17] For KK-B″, it is necessary to choose the feminine or feminine plural as the base form, since all the phonological neutralizations take place in the masculine forms.

6.1.3 Generalization to novel forms

Once a system has been learned at the problem-set level, we find it is straightforward to extend its scope to novel stems. In particular, the morpheme-parsing component can find the correct parse when given the paradigm of a novel stem by searching the parses for just that stem, with the parses for known vocabulary fixed in place; [primerəs], not in the original training set, is correctly parsed as [primer-ə-s] 'first-fem.-plur.' Given the full parsed paradigm ([prime, prime-s, primer-ə, primer-ə-s]) our system of UR-discovery readily finds the correct UR /primer/, making use of the now-fixed phonological constraint weights. We have checked this for all alternation types in all languages discussed here; a demonstration for /primer/ is given in Supplementary Materials 3-6.

6.2 Searching the whole paradigm (KK-C): Tangale phonology

The phonology of Tangale (Chadic, Nigeria) was worked out by Kidda (1993). We study a simplified data set from Kenstowicz's (1994) textbook, studied before by Cotterell et al. (2015).[18] The data consist of nouns occurring both alone and with five different suffixes.

(14) *Paradigms of Tangale*

	Target UR	*Noun*	*'the N'* /-i/	*'my N'* /-no/	*'your N'* /-go/	*'her N'* /-do/	*Gloss*
a.	/loː/	loː	loː-i	loː-no	loː-go	loː-do	'meat'
b.	/bugat/	bugat	bugat-i	bugad-no	bugat-ko	bugat-to	'window'
c.	/tugad/	tugat	tugad-i	tugad-no	tugad-go	tugad-do	'berry'
d.	/aduk/	aduk	aduk-i	adug-no	aduk-ko	aduk-to	'load'
e.	/kulug/	kuluk	kulug-i	kulug-no	kulug-go	kulug-do	'harp'
f.	/wudo/	wudo	wud-i	wud-no	wud-go	wud-do	'tooth'
g.	/taga/	taga	tag-i	tag-no	tag-go	tag-do	'shoe'
h.	/kaga/	kaga	kag-i	kag-no	kag-go	kag-do	'spoon'
i.	/jaːra/	jaːra	jaːr-i	jaːr-no	jaːr-go	jaːr-do	'arm'
j.	/ŋuli/	ŋuli	ŋul-i	ŋulno	ŋul-go	ŋul-do	'truth'
k.	/lutu/	lutu	lut-i	lut-no	lut-ko	lut-to	'bag'
l.	/duka/	duka	duk-i	duk-no	duk-ko	duk-to	'salt'

These data exemplify four phonological processes, analyzed in rule-based phonology by Kidda and Kenstowicz: (a) **Final Devoicing** of obstruents; thus (14c) /tugad/ → [tugat] 'berry'; cf. [tugad-i]. The stem in (14b), [bugat] ~ [bugat-i] illustrates a stem ending in underlyingly /t/. (b) **Syncope**: stem-final short vowels are deleted before a suffix, as in (14f) /wudo-i/ → [wudi], also /wudo-no/ → [wudno]. (c) **Progressive voicing assimilation**: Obstruent sequences undergo stem-triggered voicing assimilation; thus /bugat-go/ → [bugatko]. In rule-based analysis, this process would be fed by Syncope, as in (14k) /lutu-go/ → lutgo → [lutko]. (d) **Pre-sonorant voicing**: Obstruents placed before

[18] Following Cotterell et al., we omit tone and the phonemic ATR distinction.

a sonorant consonant are voiced, as in /bugat-no/ → [bugad-no]. Pre-sonorant voicing is opaque, being counterfed by Syncope: /lutu-no/ → [lutno]; *[ludno].

Tangale is a useful case for illustrating the KK Hierarchy. Whereas Catalan will work at multiple KK levels, Tangale requires KK-C or higher. This is because the information for finding the UR is not concentrated in any particular paradigm slot. For vowel-final stems like (14j) /ɲuli/, only the isolation form reveals the underlying identity of the final vowel, which is syncopated everywhere else. But for stems like /bugad/ that end in an obstruent, we need a suffixed form to inform us of the underlying voicing value, which is neutralized in the isolation form. In the analysis, we adopt the constraints given in (15).

(15) *Phonological constraints for Tangale*

Constraint	*Characterization*	*Fitted weight*
*V] X	Forces syncope of stem-final vowels when not word-final.	36.00
MAX(V)	Violated in cases of syncope	12.80
DEP(V)	No epenthesis (see below)	20.88
IDENT(voice) & MAX(V)	Conjoined constraint (Kirchner 1996); blocks pre-sonorant voicing in cases of syncope.	14.56
FINAL VOICED OBS.	Ban on word-final voiced obstruents	20.87
*VOICELESS BEFORE SONORANT	Forces pre-sonorant voicing	19.35
AGREE(voice)	Forces voicing agreement in suffixes	26.45
IDENT(voice)-stem	Forces the voicing agreement to follow the value of the stem consonant.	9.64
IDENT(voice)	Violated in cases of final devoicing, voicing assimilation and pre-sonorant voicing	2.95
*VTV	Used to test if the system wrongly adopts a system with intervocalic voicing	4.36[19]
REALIZEMORPH	Violated when a morpheme is not realized by any segment in the surface form; prevents /duka-i/ → *[duka] for (14l)	10.43

[19] The value of 4.36 is surprisingly high, but harmlessly so, since the countervailing constraints IDENT(voice)-stem and IDENT(voice), with much higher weights, combine to block intervocalic voicing. *VTV is not even needed for Tangale (a correct grammar can be obtained without it) but it “opportunistically” receives positive weight, since it assists Faithfulness in ruling out bad candidates like /tugad-i/ → *[tugat-i].

With the learning data of (14), the constraints of (15), and the assumption of KK-C, learning proceeded without incident, yielding the results shown in (16).

(16) *Tangale learning results*

a. *URs:* all correct URs assigned probability > 0.999
b. *Constraint weights:* as shown in (15)
c. *Surface forms:* All co1rrect forms assigned probability > 0.999

Crucially, the system chose the isolation form (e.g. /ŋuli/) as the basis for vowel stems but the contextual form (e.g. /tugad-i/) for obstruent stems, rather than consistently employing some particular paradigm slot, as KK-B″ would require. This makes it possible to derive the correct outcomes.

6.3 Segmentally-Composite URs (KK-D): Seediq phonology

Seediq (Austronesian, Taiwan) is described and analyzed in Kuo (2020, 2023), who builds on earlier work by Yang (1976). Seediq illustrates an analytical device widely employed in classical generative phonology, namely the use of “composite” URs: we find that in some paradigms, no single form provides all of the information needed to obtain a UR from which all surface forms can be derived. The problem is that a neutralizing process (here, vowel reduction) applies to every paradigm member of a stem, but not in the same location. Thus, in classical analysis the UR is “cobbled together,” taking segmental material from different allomorphs. The focus of this section is what is needed for our system to learn the classical analysis; in §7.1.2, we explore the possibility that the classical analysis is not necessarily the correct one.

Some representative data are given in (17).

(17) *Vowel reduction in Seediq*

Vowels	*Composite UR*	*no suffix*	*1 suffix*[20]	*2 suffixes*	*gloss*
Pretonic reduction:					
a. /eu/	/remux/	[ˈremux]	[r**u**ˈmux-an]	[r**u**m**u**ˈx-an-i]	‘enter’
b. /aa/	/barah/	[ˈbarah]	[b**u**ˈrah-an]	[b**u**r**u**ˈh-an-i]	‘rare’
c. /ai/	/galiq/	[ˈgaliq]	[g**u**ˈliq-an]	[g**u**l**u**ˈq-an-i]	‘break’
d. /ea/	/gedaŋ/	[ˈgedaŋ]	[g**u**ˈdaŋ-an]	[g**u**d**u**ˈŋ-an-i]	‘die’
e. /ua/	/burah/	[ˈburah]	[b**u**ˈrah-an]	[b**u**r**u**ˈh-an-i]	‘new, create’

[20] [-an] is the locative-focus present suffix; [-i] is the imperative; other suffixes yield the same phonological outcomes.

Pretonic and posttonic reduction:

f.	/ee/	/pemex/	[ˈpem**u**x]	[p**u**ˈmex-an]	[p**u**m**u**ˈx-an-i]	‘hold’
g.	/oo/	/kodoŋ/	[ˈkod**u**ŋ]	[k**u**ˈdoŋ-an]	[k**u**d**u**ˈŋ-an-i]	‘hook’

The vowel system is /i, e, a, o, u/, with stress consistently penultimate. Vowel reduction works as follows. If a vowel of any quality is *pretonic*, it is realized as [u], as in (17a-e). If a vowel is *posttonic and mid* ([e] or [o]), it is realized as [u], as in (17f-g). Since stress migrates within the paradigm, based on the number of suffixes present, different vowels of a stem will undergo reduction in different paradigm slots. The need for composite URs is demonstrated by forms like (17f), /pemex/, where we must consult the isolation allomorph to know that the UR has /e/, not /u/, as its first vowel, and the single-suffix allomorph to know that the UR has /e/, not /u/ as its second vowel. This style of analysis has been used for other languages in which stress is mobile in paradigms and stressless vowels are reduced or deleted (English: Chomsky and Halle 1968:11-12, Palauan: Flora 1974, Tonkawa: Kenstowicz and Kisseberth 1979, Russian: Crosswhite 2001, Catalan: Wheeler 2005:§2.3, Odawa: Bowers 2015, and Old Irish: Bowers 2015).

For illustrative purposes we idealized the data, leaving out forms that would illustrate additional phonological patterns discussed by Kuo.[21] Our training data consists of 15 stems (45 words), covering enough vowel patterns to test out the constraint system.

We adopt a cover constraint PENULT STRESS that enforces stress on penults and stresslessness elsewhere; it stands in for a set of appropriate constraints demanding a single word-final trochaic foot. PRETONIC REDUCTION bans all vowels but [u] in pretonic position, and POSTTONIC REDUCTION bans mid vowels in posttonic position. These constraints dominate all faithfulness constraints for vowel quality. Like Catalan, Seediq offers an instance of saltation, since posttonic /e/ is realized as [u], not the phonetically intermediate [i]. Hence we adopt the counternatural weighting of *MAP(e, i) over *MAP(e, u); we include a relatively complete set of other *MAP constraints, though the analysis would work without them.

(18) *Phonological constraints for Seediq*

Constraint	*Characterization*	*Weight*
PENULT STRESS	Stand-in for a set of metrical constraints enforcing penultimate stress.	31.18
PRETONIC REDUCTION	*any vowel except [u] when pretonic	22.09
POSTTONIC REDUCTION	*[e], [o] when posttonic	19.07

[21] Specifically: word-initial atonic vowels are dropped, vowel quality is copied regressively across [ʔ], there are various alternations in stem-final consonants, and there are also a number of irregular forms. For full discussion see Kuo (2023).

IDENT(low)	(largely redundant, given presence of *MAP(a, u) and *MAP(a, o))	1.14
IDENT(back)		4.80
IDENT(high)		5.24
IDENT(stress)	Zero-weighted (stress not phonemic)[22]	0.00
*MAP(e, i)	Forces stressless /e/ to saltate to [u].	13.73
*MAP(e, u)	Unnaturally weighted below *MAP(e, i) despite the longer phonetic path.	0.00
*MAP(a, u), *MAP(o, u), *MAP(i, u), *MAP(e, o), *MAP(a, o)	Other *MAP constraints	0.00[23] 5.00 6.09 6.03 6.22

The primary interest of the Seediq case is that it falls beyond the capacity of KK-C: for stems like (17f-g), there is no one allomorph in which all of the underlying vowels surface. Run using KK-C, the learner cannot find the classical solution, because the necessary UR candidates /pemex/ and /kodoŋ/ are not in the search space — the system ends up assigning equal probability to /pemux/ and /pumex/, with wrong surface outputs resulting. The viable URs *can* in principle be selected at KK-D, since each of their vowels does appear in at least one surface allomorph. What is needed is a way to construct appropriate UR candidates at KK-D.

6.3.1 Projecting URs at level KK-D

The idea is to use the alignments already deployed to create alternation sets (§5.3) as the basis for finding composite URs. For instance, (19a) gives the calculated optimal alignment for two surface allomorphs of (17f-g) 'hold'. The needed UR candidates can be obtained by forming all possible combinations from the choices provided in this alignment. Since there are two binary choices, we obtain four UR candidates, shown in (19b).

(19) *Forming the set of candidate URs for Seediq 'hold'*

a. Optimal string alignment

p	**e**	m	**u**	x	isolation allomorph
p	**u**	m	**e**	x	single-suffix allomorph

[22] Since stress is not phonemic in Seediq, we exclude it from the UR candidate set, following the procedure specified in §3. The GEN function of the phonology is assumed to provide candidates that include stress, as well as any other non-distinctive properties.

[23] *MAP(a, u) is weighted zero because it is largely redundant with IDENT(back), which must receive a positive weight for independent reasons.

b. Free combination of alternatives to create UR candidates

/p	e	m	u	x/	
/p	**e**	**m**	**e**	**x/**	(emerges as correct under further learning)
/p	u	m	u	x/	
/p	u	m	e	x/	

Note further that in (19) we show only the alignment of [ˈpemux] and [puˈmex], since including the third allomorph [pumux] generates no further UR candidates.

6.3.2 Learning results for Seediq

The procedure just given is all that is needed. Once a form like /pemex/ is included in the candidate set for URs, it is straightforward for our EM procedure to assign it essentially 100% probability — it is selected since it is the only one that permits correct derivation of all members of its paradigm, maximizing likelihood. The final results of the learning simulation are given in (20).

(20) *Seediq: final learning results*

a. *URs:* all correct URs assigned probability > 0.999
b. *Constraint weights:* as in (18) above
c. *Surface forms:* All correct forms assigned probability > 0.999.

One might wonder if the expanded search space made available under KK-D makes it impossible to learn Pseudo-German, Catalan, or Tangale. It turns out this is not so; in all cases KK-D returns the same UR candidate set and learning proceeds identically at both levels.

6.4 Featurally-composite URs (KK-E): "Paka-20"

The "Paka" language family was created by Tesar et al. (2003) as a heuristic data set for the study of phonological learning algorithms, including algorithms that learn URs. Since then, Paka languages have become a focus of formal work in learnability; see Alderete et al. (2005), Jarosz (2006), Merchant (2008), Tesar (2014), and Tan (2022). In light of this body of work, we see the ability to cover Paka languages as an essential criterion for proposals in UR learning, and therefore checked to make sure that our system can handle all 24 of the languages (in the version of Paka from Tesar 2014). We find that it can, but to cover *every* case necessitates ascending one more level on the KK Hierarchy.

For expository purposes it will suffice to focus here on *Paka-20* (Tesar 2014:244), which we instantiate here with the data in (21).[24]

(21) *Paka-20 paradigms*

UR	*SR*	*Gloss*	*UR*	*SR*	*Gloss*
/pa-pa/	[pápa]	‘dog-nom.’	/tʃé-pa/	[tʃépa]	‘pig-nom.’
/pa-tiː/	[páti]	‘dog-acc.’	/tʃé-tiː/	[tʃéti]	‘pig-acc.’
/pa-tʃé/	[patʃé]	‘dog-gen.’	/tʃé-tʃé/	[tʃétʃe]	‘pig-gen.’
/pa-kóː/	[pakóː]	‘dog-abl.’	/tʃé-kóː/	[tʃéko]	‘pig-abl.’
/tiː-pa/	[tíːpa]	‘cat-nom.’	/kóː-pa/	[kóːpa]	‘bat-nom.’
/tiː-tiː/	[tíːti]	‘cat-acc.’	/kóː-tiː/	[kóːti]	‘bat-acc.’
/tiː-tʃé/	[titʃé]	‘cat-gen.’	/kóː-tʃé/	[kóːtʃe]	‘bat-gen.’
/tiː-kóː/	[tikóː]	‘cat-abl.’	/kóː-kóː/	[kóːko]	‘bat-abl.’

In Paka-20, morphemes contrast for underlying stress (“accent”). The grammar assigns surface stress to the leftmost accented syllable and to the leftmost syllable in words with no underlying accent. Hence the underlyingly unaccented stems ‘dog’ and ‘cat’ appear with surface accent only in the nominative and accusative, where the suffix is unaccented; whereas underlyingly accented ‘pig’ and ‘bat’ bear invariant accent. Moreover, vowel length is phonemic, as shown by the differences between ‘dog’ and ‘cat’, ‘pig’ and ‘bat’, genitive and ablative. Underlying long vowels are shortened if they fail to receive stress, as we see for the stem vowels of ‘cat-gen./abl.’, and for the ablative suffix when it follows an accented stem.[25]

Is there a real language like Paka-20? This seems possible, but is not certain. In the analysis of Yates (2017), Cupeño (Takic, California) resembles Paka-20, but the key morphemes (with long unaccented UR) are very rare. Yates also suggests that Hittite (Indo-European, Anatolia, extinct) had an accentual pattern like Paka-20. Here, the pattern is better attested lexically, but the virtuosity of scholarly inference needed to detect phonology in the Hittite orthography makes the conclusion less certain.

For Paka-20, we follow Tesar (2014:176), with minor modifications. With the more complete set of URs and surface candidates generated by our system, we needed to add *CLASH and *LAPSE, but the change does not affect the point at hand. Our constraints are given in (22), along with the weights that were obtained in our learning simulation.

[24] Consonants and vowels are ours; glosses from Tesar.

[25] The paradigms of (21) actually give no evidence that /-tiː/ actually has an underlying long vowel; it is in effect a Rich Base form, in the sense of Prince and Smolensky (1993:209). Our system learns it with a short vowel, but weights the constraints such that hypothetical /-tiː/ would surface as desired.

(22) *Phonological constraints for Paka-20*

Constraint	*Characterization*	*Weights*
NoLong	No long vowels	0.00
Weight-to-Stress	Long vowels must be stressed.	17.64
MainLeft	Prefers leftmost stress	15.81
MainRight	Prefers rightmost stress	0.00
*Clash	Excludes doubly-stressed words	17.00
*Lapse	Excludes stressless words	24.13
Ident(long)		8.31
Ident(stress)		15.80

These weights achieve in MaxEnt the same results obtained by Tesar with classical OT. MainLeft is weighted much higher than MainRight (default choice for stress); Ident(stress) bears enough weight to overcome MainLeft (in /pa-tʃé/, [patʃé] defeats *[pátʃe]); Weight-to-Stress is weighted much higher than Ident(long) (so stressless long vowels will shorten), and Ident(long) is weighted much higher than NoLong (so that in other environments, underlying long vowels survive.)

We now explain why Paka-20 necessitates a still higher level of the KK hierarchy. The key paradigm is for ‘cat’, which requires the UR /tiː/ for its stem. This UR contains a segment, long unaccented /iː/, which never occurs in surface representations. The evidence for /iː/ in the UR comes from the stem’s phonological behavior. Its vowel must be long, because it surfaces as such when accented, contrasting with short-voweled ‘dog’. Its vowel must be unaccented because it loses out in the competition for stress to a following accented suffix, unlike ‘bat’. With URs like /tiː/, Paka-20 will defeat all versions of the KK Hierarchy thus far discussed, since they all require that a segment occur in surface representation for it to be a possible UR segment. Paka-20 thus requires us to ascend to KK-E (§3), which allows URs to contain “featurally-composite” segments.

One way to include /tiː/ in the UR candidate set for ‘cat’ — and more generally, to operationalize KK-E — is to *interpolate*: the “hidden” UR vowel /iː/ can be identified because it is phonetically intermediate between vowels found in attested allomorphs, namely /i/ and /íː/. By “intermediate,” we mean that /iː/ is [−stress] (like the [i] of /ti/), is [+long] (like the [íː] of /tíː/); and shares all of its other features with both [i] and [íː]. We offer a precise definition of “intermediate” in (23), based on similar definitions in Tesar (2014) and Magri (2018:580):

(23) *Defn.*: "*Intermediate"*

Given distinct segments x, y, z, if for every feature $\mathcal{F}$, either $\mathcal{F}(z) = \mathcal{F}(x)$ or $\mathcal{F}(z) = \mathcal{F}(y)$, we say that z is *intermediate* between x and y.[26]

This definition is used in (24), a method for forming UR candidates at KK-E:

(24) *Interpolation (KK-E)*

Let *x* and *y* be segments of two UR candidates, occupying the same column C in their aligned form ((19b)). If segment *z* is intermediate between *x* and *y*, then

1. Add *z* to column C.
2. Form the set of additional UR candidates implied by the presence of *z*.

For Paka-20, interpolation adds to the candidate UR set for 'cat' two additional UR candidates, /tiː/ and /tí/, whose vowels are intermediate between attested /íː/ and /i/. Of these, /tiː/ turns out to be correct. Interpolation also requires that we augment the method of alternation-substitution for forming surface GEN (§5.6): we must construct the alternation set using the interpolated forms along with the observed allomorphs.[27]

We omit the narrative of how our model worked here, as it is similar to previous cases; the model obtained the appropriate weights, UR probabilities, and output probabilities; for details, see Supplementary Materials 3. The key point is that cases like Paka-20 — if real cases indeed exist and are psychologically real to their speakers — can be treated by a straightforward expansion of the UR candidate set, bringing KK-E into the scope of the proposal.

Again, it is worth pondering whether the expanded search space made available under KK-E makes it impossible to learn the languages discussed earlier. We have checked and this is not so; the expansion of the candidate set is sufficiently modest that learning is unimpeded.

[26] The definition assumes fully-specified binary features, which suffice for the example under study; more would have to be said to cover underspecification (see Magri 2018:583) or multivalued features. We leave for further study the question of how intermediateness might be interpreted for segment - null pairs, such as would arise in cases of deletion or insertion; see McCarthy (2008) for relevant discussion.

[27] For example, since in Catalan [ʒ] alternates with [tʃ] (§6.1.1), surface GEN should add interpolated candidates with intermediate [ʃ] and [dʒ]. This augmentation would be needed to accommodate "saltation repair," in which language learners come to favor non-saltating candidates; see Liang et al. (2024) for an example.

6.5 Abstractness in general: what is needed to go higher?

At this point, we have completed our exploration of levels C-E of the KK Hierarchy, and turn to the implications of our KK-based system for abstractness in phonology.

We emphasize first that our system is *not averse to abstractness per se*. If some proposed abstract URs are included in the search space, and the phonological constraints needed to complete the analysis are also present, then there is no reason why our system should not be able to find a successful grammar.

We have checked this possibility for O'Hara's (2017) abstract analysis of Klamath phonology. O'Hara proposes abstract /e/ to underlie certain instances of [i] that alternate with ∅; /e/ contrasts with /i/ (which underlies non-alternating [i]) and also with ∅. Abstract /e/ never surfaces, though the phoneme /e/ is attested in other environments. None of the methods for UR candidate generation given above would posit this /e/, for it appears in no overt allomorphs, nor is it phonetically intermediate between any observed alternating segments. Thus, following O'Hara's practice, we *hand-included* /e/ in the set of candidate URs for stems with [i] ~ ∅ alternation. Given suitable constraints, our system duly adopted abstract /e/ for the relevant stems and weighted O'Hara's constraints in a way that could generate the correct outcomes; for details see Supplementary Materials 4.

However, hand-including URs in the search space skirts a serious potential difficulty. Complete models of UR learning must provide a set of hypotheses from which the correct UR may be selected. Once one has selected such a space, the issue then must be confronted of whether this space is insufficiently restrictive to be feasibly searched with the available computational resources. We turn next to this issue.

6.6 Search space matters: learning failure under a larger hypothesis space

To test the effects of a larger hypothesis space, we invented a new KK level, which we call "KK-Z." To form KK-Z, we collect all allomorphs (as in KK-C), then apply the method of alternation-substitution (§5.6). This method was used earlier to create surface GEN candidates; for KK-Z we apply it to the creation of UR candidates. For instance, because Tangale has both voicing and vowel ~ ∅ alternations, the allomorph set for 'berry' {[tugad], [tugat]} gives rise to 2^5 = 32 URs candidates; e.g. /tugad/, /tugat/, /tgad/, /dukat/, /tgd/, and so on. This is a far richer set than what would be created under KK-B″ - KK-E, all of which generate just two candidates, /tugat/ and /tugad/.

It is beyond the scope of this article to determine whether this method of forming UR candidates is applicable to actual languages.[28] We employ it here solely to show that the

[28] We are not sure whether KK-Z as such is applicable, but if it is combined with the interpolation principle of KK-E, the result is an even higher level — call it KK-EZ — which can locate the necessary URs for a number of well-known abstract analyses. For instance, KK-EZ could discover /beːn/ for English

restrictiveness of the hypothesis space can be important. To this end we returned to Tangale (§6.2), this time attempting to learn its URs using KK-Z. We discovered that the larger hypothesis space created under KK-Z led to serious learning failures, as follows: (a) aberrant voicing alternations: the (correct) UR /tugad-go/ 'your berry' is resolved 75% of the time as correct [tugadgo] and 25% of the time as *[tugatko]; (b) incorrect repair of underlying final voiced obstruents: /tugad/ surfaces as 25% correct [tugat] and 75% incorrect *[tugado], with epenthesis.

The erroneous solution can be shown to be a local optimum; it falls at log likelihood −4.49, where the log likelihood of the correct answer is 0. As external observers, we ourselves know what constraint weights should be altered, but the search algorithm cannot do this without sinking lower in log probability, since it would create greater harm among the erroneous URs still present.[29]

The situation with Catalan is similar: the grammar mishandles the [r] ~ ∅ alternation, treating cases like (11h) [kla] ~ [klar-ə] as stem allomorphy, with two equiprobable URs /klar/ and /kla/. At the same time, the system fails to learn /r/ deletion, with zero weight assigned to *CODA [r] and a high weight to MAX. The wrong empirical predictions made by this system are exemplified by the stem for 'plain', which would surface with free variation: [klar-ə]/*[kla-ə] in the feminine, and 50/50 (not the required 75/25) for [kla]/*[klar] in the masculine.

The results just given are based on the initial parameter values of §6. However, we obtain the same outcomes under a different plausible initialization (Tesar and Smolensky 2000) in which Markedness constraints start out high (50) and Faithfulness low (1).

It is worth trying to diagnose these learning failures. We suggest that when our system is confronted with a very large set of URs, as with KK-Z, it is liable to embark on the spurious task of making sure that the more remote URs map onto correct surface forms. This makes strong demands on the setting of phonological weights, demands that turn out to be not addressable by the learning mechanism. The learner arrives at local maxima because it is overburdened with "imaginary" derivations that under more

---

bean [biːn] (*SPE* p. 69), since /eː/ is intermediate between [iː] and [ɛ], which alternate in *serene* ~ *serenity* (*SPE* p. 55). Yowlumne /uː/ for [oː] could also be discovered, since in some cases [oː] alternates with [u] (Kenstowicz and Kisseberth 1977: 50). Dida /A/ ([+low, +ATR]) for [e] could be discovered since it is intermediate between [e] and [a], which alternate (Kaye 1980:8). We have not tried running learning simulations using KK-EZ, since for the point made in this section, KK-Z suffices.

[29] If we hand-enter correct $\boldsymbol{\theta}$ values for the UR candidates created under KK-Z, our system discovers correct constraint weights, and if we hand-enter the correct constraint weights, the system discovers correct $\boldsymbol{\theta}$ values. It is the mutual dependence of $\boldsymbol{\theta}$ and $\boldsymbol{W}$ that leads the system astray.

conservative KK levels simply do not arise. Once the system has settled at a local optimum, it is defeated, frozen in place by the need to solve pseudo-problems.[30]

To restate our findings more generally, we have set up, in effect, a controlled experiment in which a consistent system of calculation is applied to a sequence of ever-larger hypothesis spaces (KK-B″, …, KK-Z). We show that at the last step, the hypothesis space has taken on a form that is too difficult to search in the cases of Tangale and Catalan. We turn to our interpretation of this result in the following sections.

7. Discussion

7.1 Relation to abstractness

We have shown that our proposed learning system, when provided with a phonological constraint set, can learn morphophonemic systems of the scope and complexity seen in ordinary problem sets. The key ingredients are (a) early detection of morpheme membership by similarity, so that allomorph sets can be found; (b) use of alignment to locate alternating segments; (c) algorithms that project UR candidate sets at each level of the KK Hierarchy; and (d) Expectation-Maximization for finding the right UR choices and constraint weights.

We used this system to address the abstractness controversy by varying the level of the KK Hierarchy that provides the candidate UR set. The key result is that our system works smoothly at lower KK levels (up to E), but runs aground on local maxima when we scale up to KK-Z. We offer this as a demonstration that the old controversy about abstractness can in principle be addressed in more explicit terms by using a computational learning model.

7.1.1 The criterion of success

The key context of our demonstration is the idea that the right goal for computational learning is not necessarily to achieve the best-performing system — for example, a model that really can learn URs correctly when the UR hypothesis space is set at KK-Z. Rather, adopting the Kiparskian perspective from §2, we seek to devise a system that succeeds where young humans succeed, and fails — carries out restructuring — where young humans do the same. The UR abstractness controversy is founded on this point.

[30] As reviewers point out, some of the difficult cases might be fixed by invoking further principles of phonological theory. For Tangale, if Universal Grammar requires that DEP be much more highly weighted than IDENT(voice), then the bad candidate *[tagado] from /tugad/ might be ruled out. We have not found any comparable way to avoid the other errors mentioned above. The discussion above offers a principled reason to expect that KK-Z will, applied to multiple languages, repeatedly get stuck in comparable local maxima.

7.1.2 What is the right KK-level?

The goal of predicting restructuring is applicable to multiple areas, including opacity (Kiparsky 1971) and structural complexity (Moreton and Pater 2012). For the specific case of UR abstractness, we frame the question as follows: is there some KK level above which we can correctly predict that restructuring will occur?

In the present stage of research, it is premature to offer a firm answer. Were we to propose a particular KK level most likely to be correct, our guess would be KK-C or lower. This guess is based on increasing empirical evidence indicating that restructuring has occurred in languages for which the traditional analysis of the data pattern requires KK-D or higher, thus throwing the validity of analysis at this level (and by implication, higher levels) into doubt. Such work includes studies of Yidiɲ (Hayes 1999); Lakhota (Albright 2002); Old Irish and Russian (Bowers 2015); Nishnaabemwin (Bowers 2019), and, in fact, Seediq, which we discuss here.

7.1.3 Choosing a KK-level for phonological theory: the Seediq evidence

The classical analysis for Seediq that our system learns (§6.3) derives from the early (post-*SPE*) study of Yang (1976). However, our primary source, Kuo (2020, 2023), making use of additional data and experimental results, argues that this analysis is actually incorrect. While it does accurately recapitulate the sequence of sound changes that Seediq underwent (penultimate stress, followed by vowel reduction), it does not correctly characterize the internalized grammar of present-day Seediq speakers, who in fact have adopted a restructured phonology, detectible in diachronic changes for a number of stems. As Kuo shows, these are the stems (like (17f-g)) that would require KK-D-type URs under the analysis put forth above. The restructuring worked as follows: the UR is now identical to the isolation form, and the "restored" vowel that appears when stress is shifted does not come from this UR, but is simply a *copy of the penultimate stem vowel*; e.g. [ˈp**e**mux] ~ [puˈm**e**x-an]. This is an instance of the "prosodic correspondence" discovered by Crosswhite (1998). The vowel-copying generalization was already present on a statistical basis even before vowel reduction entered the language; the later restructuring consisted of extending vowel copying productively to additional stems. Still further restructuring was discovered in the wug-test experiment reported by Kuo (2023): the participants extended vowel copying even to the vowel [a], creating alternations like [ˈa u] ~ [u ˈa - an] that do not occur in existing Seediq words.

Kuo's own modeling studies show that the diachronic changes and wug-test results can be accounted for by a MaxEnt learning model using KK-B″ for its underlying representations. In this section, we focus instead on the difference between KK-C and KK-D; our results would carry over straightforwardly to KK-B″. We redid our Seediq learning simulations at both KK-C and KK-D, this time including Kuo's constraint (2023:5) NUC-IDENT-OO[F], which is responsible for vowel copying. We also use a more realistic training set, with sufficient stems to represent the preference for vowel copying among existing forms. We added a small number of hypothetical forms to represent stem types like / u e / that originally existed in Seediq but have been ironed out by restructuring. Thus the training set was expanded to include all 25 logically possible stem types; frequencies were as in the Kuo corpus, but with frequency 1 for hypothetical forms. When we ran our learning system on these data at KK-D, we found that its performance was, in essence, "too good", a very close match to the training data. NUC-IDENT-OO[F] received a near-zero weight, because under KK-D it is superfluous, its work being done instead by the KK-D-style URs. Hence, the language remained completely stable — descriptive success, but explanatory failure. In contrast, at KK-C our system assigned NUC-IDENT-OO[F] a substantial weight, picking up on the modest pre-existing tendency toward vowel-matching. As a result, the learned KK-C grammar generated "restructured" forms similar to those documented by Kuo in historical change and in her wug-test; for example, the KK-C grammar strongly preferred vowel-copied [u ˈe - an] as a suffixed version of [ˈe u]. Full details of these simulations are provided in Supplementary Materials 3.4.

In sum, if we are to explain why a new generation of Seediq speakers restructured their phonology, a plausible basis is to assume that the KK-D level analysis was not accessible to them. The additional cases cited above provide further evidence that phonological theory plausibly might adopt KK-C (or lower) as the upper limit of the hypothesis space for URs.[31]

7.1.4 Informativeness and KK-B″

We have said little so far about KK-B″, the single surface-base hypothesis. Albright's proposal (§3) is that in early morphophonemic learning, children try out a variety of paradigm slots to serve as the basis for the UR, ultimately settling on the most effective one and using it exclusively thereafter. We suggest that this selection process might effectively be modeled at the level of KK-C, where multiple stem allomorphs compete to serve as UR. Once a sufficient number of cases are worked out, it would be straightforward to determine which paradigm slot most often provides a feasible UR (this would be the plural for Pseudo-German, the feminine for Catalan, the isolation form for Seediq, and so on). Once this choice is made, the learner would subsequently rely on this

[31] Reviewers and colleagues have suggested to us an alternative approach in which the child tends to stick to one KK level, but can entertain higher levels when abundant data support such a move. This idea is already adumbrated by KK's notion (1977:4) of learning "under duress." The idea is plausible, but for the empirical cases discussed in this section it requires special pleading: e.g. for Seediq we would need to assume that the learning data were insufficiently abundant to justify the formation of a KK-D analysis, and similarly for the other cases mentioned above.

choice for future learning. Thus, our approach is not incompatible with KK-B″, but offers a way to implement it.

7.2 Further issues

7.2.1 The need for paradigms

Our system learns URs by comparing paradigm members with each other; this can only happen if a sufficient fraction of the paradigm members is co-present for processing in the child's mind/brain. This premise can be assessed empirically, since there are experimental probes that inform us which words are listed in the lexicon; for discussion and literature review see Baayen et al. (2002). Such work suggests that the lexicon includes not just listed irregulars, but many regular forms as well. The only regulars that give evidence of *not* being listed are those of lower frequency. This pattern suggests that children do, at the first stages, memorize a great many paradigms which could serve them for learning, as in the scheme we and others have adopted.

It is not necessary to know *complete* paradigms, such as we have used, in order for learning to proceed. Experimenting with "gappy" learning data, we find that our system can cope with accidental gaps — to be sure, some gaps inevitably lead to wrong UR guesses (e.g. */gris/, not /griz/ for (11k), Catalan [gris], if the gaps are in the feminine), but the phonological constraint weighting is nonetheless learned correctly if sufficient data are present to justify each phonological process.

This said, it will ultimately be necessary to discover how the input data get organized into paradigms in the first place — how do children come to know that *jumping* is an inflected form of *jump*? Perhaps phonetic similarity, assisted by meaning, helps the child in grouping allomorphs into common morphemes. It is encouraging that humans appear to be able to detect at least the *presence* of affixes even in infancy, starting at six months (Kim and Sundara 2021). Humans must also apprehend what morphological categories are present. For some first efforts to address these tasks, see Baroni et al. (2002), Dreyer and Eisner (2011), and Jin et al. (2020).

7.2.2 Projecting paradigm members from partial information

A longer-term goal for research such as ours is to integrate computational learning models into broader phonological research, providing explicit hypotheses that can be tested with typological study, acquisition data, and experiments. For instance, an adequate model might be expected to make correct predictions about the outcome of a wug test. Already, proposed learning models have been tested against wug-test data (e.g., Albright and Hayes 2003, Ernestus and Baayen 2003), but such tests have often employed models that do not make use of URs.

The reason our system would not do well on a wug test is that, like other UR-based models, it assumes optimum learning conditions; it is presented with a complete paradigm for a morpheme and asked to infer the UR from it. But in a wug test — or indeed, in real life — the language user often possesses *incomplete* information and must use it to synthesize a novel form. Ernestus and Baayen's 2003 landmark study revealed that speakers can do well under such circumstances; specifically, they make use of multiple stochastic cues for the purpose of UR-guessing and deploy them to take the wug test. In Dutch, Ernestus and Baayen's focus, the isolation form is not optimal for guessing URs, since the language has Final Devoicing, but Dutch speakers can guess the UR of an isolation form far better than chance by relying on cues based on the place and manner of the word-final obstruent, as well as the length of the preceding vowel. Similar results have been found in other languages; for references see the discussion of frequency-matching in §6.1.1. The upshot is that, to wug-test well, a complete system will have to go beyond UR learning with complete information, incorporating further principles for rational guessing under incomplete information.

7.2.3 Allomorphy, irregularity, inflectional classes

Our system can solve phonology problem sets, which normally do not include any patterns not attributable to productive phonology. These include listed affix allomorphs (like Korean nominative *-ka*/*-i*), irregular stems (e.g. the [kɛp] of *kept*), suppletion (*go/went*), and conjugation/ declension classes. We suggest that all of these phenomena must be treated in a module for morphology. The phonology of alternations, in one sense, is the child's method for assigning greater systematicity to the morphological pattern, often permitting her to synthesize novel allomorphs (Kenstowicz and Kisseberth 1979:46-55). By this view, an adequate theory of morphophonemic learning must embed the phonological part into a larger theory that handles aspects of morphology that are not phonologically reducible.

7.2.4 Learning the constraints

A more ambitious version of our learner would use its own constraints rather hand-provided ones. An obvious source for learned Markedness constraints is phonotactic learning. Already, many proposals have been made to induce constraints on the basis of learning data (Hayes and Wilson 2008, Heinz 2010, Jardine and Heinz 2016, Wilson and Gallagher 2018, Gouskova and Gallagher 2020, Dai 2021, Rawski 2021, Hua and Jardine 2021, Kim 2022). This strategy forges a learning-theoretic connection between phonotactics and alternation, in which the former yields Markedness constraints employed in the latter. For Faithfulness, observe that when alternation-substitution GEN (§5.6) is adopted, the task of finding an adequate constraint set becomes easier, because there will only be a modest number of unfaithful candidates, far fewer than in the "freedom of analysis" approach of standard OT (McCarthy 2007), which aspires also to cover phonotactics. With alternation-substitution GEN, it likely suffices to select from the finite schemata of MAX, DEP, IDENT, *MAP, and so on, for sufficient constraints to cover the candidate set, which itself is limited in size.

As elsewhere, it is opacity that creates the biggest challenges. For instance, the constraint IDENT(voice) & MAX(V) we used for Tangale opacity would not be included in the simple schemata enumerated above. Opacated Markedness constraints, like *FINAL [n] and *CODA [r] in Catalan, have surface counterexamples in derivations like /sant/ → [san]; these would likewise need additional mechanisms or grammar architectures (Nazarov and Pater 2017).

7.2.5 More powerful learners?

We are curious whether a computational system could be devised that, unlike ours, could learn effectively at high, abstractness-friendly KK levels. For instance, we noted (fn. 28) that a level we called "KK-EZ" provides a hypothesis space sufficient to include the URs of many well-known highly abstract analyses. A learning model that was able to navigate such spaces successfully, avoiding local maxima, would in our opinion be a meaningful contribution to the abstractness debate.

References

Albright, Adam. 2008. A restricted model of UR discovery: Evidence from Lakhota. Ms., MIT, Cambridge, MA.

Albright, Adam. 2005. The morphological basis of paradigm leveling. In *Paradigms in phonological theory*, ed. by Laura J. Downing, Tracy Alan Hall, and Renate Raffelsiefen, 17–43. London: Oxford University Press.

Albright, Adam. 2008. Explaining universal tendencies and language particulars in analogical change. In *Language universals and language change*, ed. by Jeff Good, 144–181. London: Oxford University Press.

Albright, Adam. 2010. Base-driven leveling in Yiddish verb paradigms. *Natural Language and Linguistic Theory* 28:475-537.

Adam Albright, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: a computational/experimental Study. *Cognition* 90:119-161.

Albro, Daniel M. 2005. Computational Optimality Theory and the phonological system of Malagasy. Doctoral dissertation, UCLA.

Alderete, John, Adrian Brasoveanu, Nazarré Merchant, Alan Prince, and Bruce Tesar. 2005. Contrast analysis aids the learning of phonological underlying forms. In WCCFL 24, ed. by John Alderete, Chung-hye Han and Alexei Kochetov, 34-42. Somerville, MA: Cascadilla Press.

Apoussidou, Diana. 2007. The learnability of metrical phonology. Doctoral dissertation, University of Amsterdam.

Archangeli, Diana and Douglas Pulleyblank. 1994. Grounded phonology. Cambridge, MA: MIT Press.

Aziza, Rose O. 2008. Neutralization of contrast in the vowel system of Urhobo. Studies in African Linguistics 37:1-19.

Baayen, Harald, Robert Schreuder, Nivja de Jong, and Andrea Krott. 2002. Dutch inflection: the rules that prove the exception. In Storage and computation in the language faculty, ed. by Sieb Nooteboom, Fred Weerman, and Frank Wijnen, 61-92. Dordrecht: Springer.

Bailey, Todd M. and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? Journal of Memory and Language 44:568–591.

Baković, Eric. 2009. Abstractness and motivation in phonological theory. Studies in Hispanic and Lusophone Linguistics, 2:183-198.

Baković, Eric, Jeffrey Heinz, and Jonathan Rawski. 2022. Phonological abstraction in the mental lexicon. In Oxford Handbook of the Mental Lexicon, ed. by Anna Papafragou, John Trueswell and Lila Gleitman, 11-32. Oxford University Press.

Baroni, Marco, Johannes Matiasek, and Harald Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning, 48–57. Association for Computational Linguistics.

Barke, Shraddha, Rose Kunkel, Nadia Polikarpova, Eric Meinhardt, Eric Bakovic, and Leon Bergen. 2019. Constraint-based Learning of Phonological Processes. In EMNLP-IJCNLP 2019, 6176–6186. Association for Computational Linguistics.

Becker, Michael, Andrew Nevins and Jonathan Levine. 2012. Asymmetries in generalizing alternations to and from initial syllables. Language 88:231-268.

Belth, Caleb. 2023. Towards a learning-based account of underlying forms: A case study in Turkish. Proceedings of the Society for Computation in Linguistics, 2023.

Bonet, Eulalia and Maria-Rosa Lloret. 2018. Fricative-affricate alternations in Catalan. Probus 30:215-249.

Bowers, Dustin. 2015. A system for morphophonological learning and its consequences for language change. Doctoral dissertation, UCLA.

Bowers, Dustin. 2019. The Nishnaabemwin [Odawa] restructuring controversy: New empirical evidence. Phonology 36:187-224.

Brame, Michael. 1972. On the abstractness of phonology: Maltese ʕ. In Contributions to generative phonology, ed. by Michael Brame, 22-61. Austin: University of Texas Press.

Bynon, Theodora. 1977. Historical Linguistics. Cambridge: Cambridge University Press.

Calamaro, Shira and Gaja Jarosz. 2015. Learning general phonological rules from distributional information: A computational model. Cognitive Science 39: 647–666

Chomsky, Noam and Morris Halle. 1968. The Sound Pattern of English. New York: Harper and Row.

Chong, Adam. 2019. Exceptionality and derived-environment effects: A comparison of Korean and Turkish. Phonology 36:543-572.

Chong, Adam. 2021. The effect of phonotactics on alternation learning. Language 97: 213-244.

Comrie, Bernard. 1986. The Maltese pharyngeal. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 39:12-18.

Cotterell, Ryan, Nanyun Peng and Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. Transactions of the Association for Computational Linguistics, 3:433-447.

Crosswhite, Katherine. 1998. Segmental vs. prosodic correspondence in Chamorro. Phonology 15:281-316.
Crosswhite, Katherine. 2001. Vowel reduction in Optimality Theory. New York: Routledge.
Czaykowska-Higgins, Ewa. 1988. Investigations into Polish morphology and phonology. Doctoral dissertation. MIT.
Dai, Huteng. 2021. Learning nonlocal phonotactics in a strictly piecewise probabilistic phonotactic model. In AMP 2021, 301-203. Linguistic Society of America.
Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19:380–393.
Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39:1-22.
Derwing, Bruce L. 1973. Transformational grammar as a theory of language acquisition. Cambridge: Cambridge University Press.
Dowd, Andrew. 2005. Surface base selection in Pengo. In WCCFL 24, ed. by John Alderete, Chung-hye Han and Alexei Kochetov, 105-111. Somerville, MA: Cascadilla Proceedings Project.
Dresher, Bezalel Elan 1981. On the learnability of abstract phonology. In The logical problem of language acquisition, ed. by C. L. Baker and John J. McCarthy, 188–210. Cambridge: MIT Press.
Dreyer, Markus and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In EMNLP 2011, 616–627. Edinburgh: Association for Computational Linguistics.
Eisner, Jason. 1997. Efficient generation in primitive Optimality Theory. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 313–320. East Stroudsburg, PA: Association for Computational Linguistics.
Eisenstat, Sarah. 2009. Learning underlying forms with MaxEnt. Master’s thesis, Brown University.
Ellis, Kevin, Adam Albright, Armando Solar-Lezama, Joshua B. Tenenbaum, and Timothy J. O’Donnell. 2022. Synthesizing theories of human language with Bayesian program induction. Nature Communications 13:5024.
Ernestus, Mirjam, and R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. Language 79:5-38.
Flora, Marie Jo-Ann. 1974. Palauan phonology and morphology. Doctoral dissertation, University of California, San Diego. San Diego, CA.
Foley, James. 1965. Spanish morphology. Doctoral dissertation, MIT, Cambridge, MA.
Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Proceedings of the Stockholm Workshop on Variation within Optimality Theory, ed. by Jennifer Spenader, Anders Eriksson and Östen Dahl, 111–120. Stockholm: Stockholm University.
Gorecka, Alicja. 1988. Epenthesis and the coda constraints in Polish. Ms., MIT.
Gussmann, Edmund. 1980. Studies in abstract phonology. Cambridge, MA: MIT Press.

Gouskova, Maria. 2012. Unexceptional segments. Natural Language and Linguistic Theory 30:79-133.

Gouskova, Maria and Gillian Gallagher. 2020. Inducing nonlocal constraints from baseline phonotactics. Natural Language and Linguistic Theory: 38:77-116.

Hammarström, Harald and Borin, Lars. 2011. Unsupervised learning of morphology. Computational Linguistics 37:309-350.

Hansson, Gunnar O., and Ronald Sprouse. 1999. Factors of change: Yowlumne vowel harmony then and now. In Proceedings of WSCLA IV, 39-57.

Hayes, Bruce. 1999. Phonological restructuring in Yidiny and its theoretical consequences. In The derivational residue in phonological Optimality Theory, ed. by Ben Hermans and Marc Oostendorp, 175-205. Amsterdam: John Benjamins.

Hayes, Bruce. 2008. Introductory phonology. Oxford: Blackwell.

Hayes, Bruce and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry 39: 379–440.

Hayes, Bruce and James White. 2015. Saltation and the P-map. Phonology 32:267-302.

Hayes, Bruce, Kie Zuraw, Peter Siptar, and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. Language 85:822-863.

Heinz, Jeffrey. 2010. Learning long-distance phonotactics. Linguistic Inquiry 41:623–661.

Hua, Wenyue, and Adam Jardine. 2021. Learning input strictly local functions from their composition. In Proceedings of the Fifteenth International Conference on Grammatical Inference, 47-65.

Hoffman, Carl. 1973. The vowel harmony system of the Okpe monosyllabic verb or Okpe: A nine-vowel language with only Seven vowels. Research Notes from the Department of Linguistics and Nigerian Languages 6:79-111. University of Ibadan.

Hooper, Joan B. 1976. An introduction to natural generative phonology. New York: Academic Press.

Hudson Kam, Carla L., and Elissa L. Newport. 2009. Getting it right by getting it wrong: When learners change languages. Cognitive Psychology 59:30-66.

Ito, Junko, and Armin Mester. 2003. On the sources of opacity in OT: Coda processes in German. In The syllable in Optimality Theory, ed. by Caroline Féry and Ruben van de Vijver, 271-303. Cambridge University Press.

Jardine, Adam, and Jeffrey Heinz. 2016. Learning tier-based strictly 2-local languages. Transactions of the Association for Computational Linguistics 4:87–98.

Jarosz, Gaja. 2006. Rich lexicons and restrictive grammars: Maximum likelihood learning in Optimality Theory. Doctoral dissertation, Johns Hopkins University.

Jarosz, Gaja. 2008. Partial ranking and alternating vowels in Polish. In Proceedings of the Chicago Linguistic Society 41, ed. by Rodney L. Edwards, Patrick J. Midtlyng, Colin L. Sprague, and Kjersti G. Stensrud, 193-206. Chicago: CLS.

Jarosz, Gaja. 2013. Learning with hidden structure in Optimality Theory and Harmonic Grammar: Beyond robust interpretive parsing. Phonology 30: 27-71.

Jarosz, Gaja. 2019. Computational modeling of phonological learning. Annual Review of Linguistics 5:67-90.

Jin, Huiming, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6696–6707. Association for Computational Linguistics.

Johnson, Mark, Joe Pater, Robert Staubs, and Emmanuel Dupoux. 2015. Sign constraints on feature weights improve a joint model of word segmentation and phonology. In Proceedings of NAACL 2015, 303–313. Denver: Association for Computational Linguistics.

Jusczyk, Peter W., Angela D. Friederici, Jeanine M.I. Wessels, Vigdis Y. Svenkerud and Ann Marie Jusczyk. 1993. Infants' sensitivity to the sound patterns of native language words. Journal of Memory and Language 32:402-420.

Kaye, Jonathan. 1980. The mystery of the tenth vowel. Journal of Linguistic Research 1:1-14.

Kang, Yoonjung. 2006. Neutralization and variations in Korean verbal paradigms. Harvard Studies in Korean Linguistics 11:183–196.

Kenstowicz, Michael. 1994. Phonology in generative grammar. Cambridge, MA: Blackwell.

Kenstowicz, Michael and Charles Kisseberth. 1977. Topics in phonological theory. New York: Academic Press.

Kenstowicz, Michael, and Charles Kisseberth. 1979. Generative phonology: Description and theory. New York: Academic Press.

Kidda, Mairo E. 1993. Tangale phonology: A descriptive analysis. Berlin: Reimer.

Kim, Seoyoung. 2022. Restrictive tier induction. Doctoral dissertation, University of Massachusetts Amherst.

Kim, Yun Jung, and Megha Sundara. 2021. 6-month-olds are sensitive to English morphology. Developmental Science, e13089.

Kiparsky, Paul. 1968. Linguistic universals and linguistic change. In Universals in linguistic theory, ed. by Emmon Bach and Robert T. Harms, 170-202. New York: Holt, Rinehart & Winston.

Kiparsky, Paul. 1971. Historical linguistics. In A survey of linguistic science, ed. by William Orr Dingwall, 576-642. College Park: University of Maryland Linguistics Program.

Kiparsky, Paul. 1973. Abstractness, opacity, and global rules. In Three dimensions in linguistic theory, ed. by Osamu Fujimura, 57–86. Tokyo: TEC.

Kiparsky, Paul. 1982. Lexical phonology and morphology. In Linguistics in the Morning Calm, ed. by I.-S. Yang, 3-91. Seoul: Hanshin.

Kirchner, Robert. 1996. Synchronic chain shifts in Optimality Theory. Linguistic Inquiry 27: 341-350.

Khalifa, Salam, Sarah Payne, Jordan Kodner, Ellen Broselow, and Owen Rambow. 2023. A cautious generalization goes a long way: Learning morphophonological rules. In ACL 2023, 1793-1805. Toronto: Association for Computational Linguistics.

Kruskal, Joseph. 1983. An overview of sequence comparison: Time warps, string edits, and macromolecules. SIAM Review 25:201-237.

Kuo, Jennifer. 2020. Evidence for base-driven alternation in Tgdaya Seediq. M.A. thesis. UCLA.
Kuo, Jennifer. 2023. Evidence for prosodic correspondence in the vowel alternations of Tgdaya Seediq. Phonological Data and Analysis 5:1-31.
Kuo, Jennifer. 2024. Phonological reanalysis is guided is guided by markedness: The case of Malagasy weak stems. Phonology 41:e3.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. Language 45:715–62.
Ladányi, Enikő, Ágnes M. Kovács, and Judit Gervain. 2020. How do 15-month-old infants process morphologically complex forms in an agglutinative language? Infancy 25:190–204.
Liang, Kevin, Victoria Mateu, and Bruce Hayes. 2024. A wug-test study of the Catalan consonant alternations. Paper given at the Annual Meeting on Phonology, Rutgers University.
Lightner, Theodore. 1965. Segmental phonology of modern standard Russian. Doctoral dissertation, MIT.
Lubowicz, Anna. 2002. Derived environment effects in Optimality Theory. Lingua 112:243–280.
McCarthy, John J. 2007. What is Optimality Theory? Language and Linguistics Compass 93:260-291.
McCarthy, John J. 2008. The gradual path to cluster simplification. Phonology 25:271-319.
McCarthy, John, and Alan Prince. 1995. Faithfulness and reduplicative identity. In University of Massachusetts occasional papers 18: Papers in Optimality Theory, 249–384. GLSA, University of Massachusetts, Amherst.
McLachlan, Geoffery J., and David Peel. 2000. Finite mixture models. New York: Wiley.
Magri, Giorgio. 2018. Output-drivenness and partial phonological features. Linguistic Inquiry 49:577–598.
Marquis, Alexandra and Shi, Rushen. 2012. Initial morphological learning in preverbal infants. Cognition 122:61–66.
Mascaró, Joan. 1976. Catalan phonology and the phonological cycle. Doctoral dissertation, MIT.
Merchant, Nazarré. 2008. Discovering underlying forms: Contrast pairs and ranking. Doctoral dissertation, Rutgers University.
Moore-Cantwell, Claire, and Robert Staubs. 2014. Modeling morphological subgeneralizations. In Proceedings of Phonology 2013, ed. by John Kingston, Claire Moore-Cantwell, Joe Pater, and Robert Staubs. Linguistic Society of America. Available at http://dx.doi.org/10.3765/amp.v1i1.42.
Moreton, Elliott, and Joe Pater. 2012. Structure and substance in artificial-phonology learning. Part I, Structure. Language and Linguistics Compass 6:686-701.
Nazarov, Aleksei, and Joe Pater. 2017. Learning opacity in stratal maximum entropy grammar. Phonology 34:299-324.
Nelson, Max. 2019. Segmentation and UR acquisition with UR constraints. In Proceedings of the Society for Computation in Linguistics: Vol. 2, Article 8.

Newman, Stanley. 1944. Yokuts Language of California. New York: Viking Fund.
Nyman, Alexandra. 2021. Learnability of a phonetically null segment. *University of Pennsylvania Working Papers in Linguistics* 27: Issue 1, Article 21.
Nyman, Alexandra, and Bruce Tesar. 2019. Determining underlying presence in the learning of grammars that allow insertion and deletion. Glossa 4:1-41.
Odden, David. 2005. Introducing phonology. Cambridge: Cambridge University Press.
O'Hara, Charlie. 2017. How abstract is more abstract? Learning abstract underlying representations. Phonology 34:325-345.
Paster, Mary. 2013. Rethinking the 'duplication problem'. Lingua 126:78-91.
Pater, Joe, Karen Jesney, Robert Staubs, and Brian Smith. 2012. Learning probabilities over underlying representations. In Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, ed. by Adam Albright and Lynne Cahill, 62–71. Montreal: Association for Computational Linguistics.
Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. Cognition 101: B31-B41.
Prince, Alan, and Paul Smolensky. 1993/2004. Optimality Theory: Constraint interaction in generative grammar. Technical report, Rutgers University Center for Cognitive Science. [Published 2004; Oxford: Blackwell].
Rasin, Ezer, and Roni Katzir. 2016. On evaluation metrics in optimality theory. Linguistic Inquiry 47:235-282.
Rasin, Ezer, and Roni Katzir. 2018. Learning abstract underlying representations from distributional evidence. In Proceedings of NELS 48, ed. by Sherry Hucklebridge and Max Nelson. 283-290. GLSA, University of Massachusetts, Amherst.
Rasin, Ezer, and Roni Katzir. 2020. A conditional learnability argument for constraints on underlying representations. Journal of Linguistics 56:745-773.
Rasin, Ezer, Iddo Berger, Nur Lan, Itamar Shefi, and Roni Katzir. 2021. Approaching explanatory adequacy in phonology using Minimum Description Length. Journal of language modelling 9:17-66.
Rawski, Jonathan. 2021. Structure and learning in natural language. Doctoral dissertation, Stony Brook University.
Richter, Caitlin. 2021. Alternation-sensitive phoneme learning: Implications for children's development and language change. Doctoral dissertation, University of Pennsylvania.
Riggle, Jason. 2004. Generation, recognition, and learning in finite state Optimality Theory. Doctoral dissertation, UCLA.
Rubach, Jerzy. 1984. Cyclic and lexical phonology: The structure of Polish. Dordrecht: Foris.
Rysling, Amanda. 2016. Polish yers revisited. Catalan Journal of Linguistics 15:121-143.
Schumacher, R. Alexander, and Janet B. Pierrehumbert. 2021. Familiarity, consistency, and systematizing in morphology. Cognition 212:104512.
Sommerstein, Alan. 1977. Modern phonology. London: Edward Arnold.

Steriade, Donca. 2000. Paradigm uniformity and the phonetics-phonology boundary. In Papers in laboratory phonology V, ed. by Michael B. Broe and Janet B. Pierrehumbert, 313–334. Cambridge: Cambridge University Press.

Sundara, Megha, James White, Yun Jung Kim and Adam Chong. 2021. Stem similarity modulates infants' acquisition of phonological alternations. *Cognition* 209:104573.

Tan, Adeline. 2022. Concurrent hidden structure & grammar learning. In Proceedings of the Society for Computation in Linguistics: Vol. 5, Article 5.

Tesar, Bruce. 2014. Output-driven phonology: Theory and learning. Cambridge: Cambridge University Press.

Tesar, Bruce, and Paul Smolensky. 2000. Learnability in Optimality Theory. Cambridge: MIT Press.

Tesar, Bruce, John Alderete, Graham Horwood, Nazarre´ Merchant, Koichi Nishitani, and Alan Prince. 2003. Surgery in language learning. In WCCFL 22: Proceedings of the 22nd West Coast Conference on Formal Linguistics, ed. by Gina Garding and Mimu Tsujimura, 477–490. Somerville, MA: Cascadilla Proceedings Project.

Tranel, Bernard. 1981. Concreteness in generative phonology: Evidence from French. Berkeley: University of California Press.

Vago, Robert. 1976. Theoretical implications of Hungarian vowel harmony. Linguistic Inquiry 7:243-263.

Wheeler, Max. 2005. The Phonology of Catalan. Oxford: Oxford University Press.

White, James. 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. Language 93:1-36.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. Cognitive Science 30:945-982.

Wilson, Colin, and Gillian Gallagher. 2018. Accidental gaps and surface-based phonotactic learning: a case study of South Bolivian Quechua. *Linguistic Inquiry* 49:610-623.

Yang, Hsiu-fang. 1976. The phonological structure of the Paran dialect of Sediq. [In Chinese] Bulletin of the Institute of History and Philology, Academia Sinica 47:611-706.

Yates, Anthony. 2017. Lexical accent in Cupeño, Hittite, and Indo-European. Doctoral dissertation, UCLA.

Zuraw, Kie. 2000. Patterned exceptions in phonology. Doctoral dissertation, UCLA.

Zuraw, Kie. 2007. The role of phonetic knowledge in phonological patterning: corpus and survey evidence from Tagalog. Language 83:277–316.

Zuraw, Kie. 2013. *MAP constraints. Ms, UCLA. Available at www.linguistics.ucla.edu/people/zuraw/dnldpprs/star_map.pdf.