Learning phonological underlying representations: the role of abstractness*

Yang Wang
University of Utah

Bruce Hayes
UCLA

To appear in *Linguistic Inquiry*

This version is the final version from the authors, submitted to *LI* for copy-editing.

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 of Kiparsky and others 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*, Arto Anttila, Eric Baković, Jane Chandlee, Tim Hunter, Adam Jardine, Gaja Jarosz, Claire Moore-Cantwell, Charlie O'Hara, Paul Kiparsky, Adeline Tan, Colin Wilson, Tony Yates, the members of the UCLA Phonology Seminar, and audiences at 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 phonological grammar that maps them into surface forms. Characteristically, researchers have employed training sets consisting of morphological paradigms in phonetic transcription, and developed systems that can recover the underlying representation (UR) for each morpheme present, along with the phonological grammar whereby the surface forms can be derived.

This has proven to be an interesting and challenging task, and it has been undertaken with a wide variety of approaches: error-driven learning with ranked constraints (e.g. Tesar et al. 2003; Apoussidou 2006, 2007; Merchant 2008; Merchant and Tesar 2008; Tesar 2014; Nyman and Tesar 2019); distributional learning under the principle of 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 et al. 2020, Hua and Jardine 2021); Bayesian Program Synthesis with ordered rules (Ellis et al. 2015, Barke et al. 2019, Ellis et al. 2022), algorithmic learning approaches guided by different evaluation criteria (Khalifa et al. 2023, Belth 2023), and maximum likelihood learning under different probabilistic frameworks (Jarosz 2006, 2015; 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 existing research tradition (see §5). We differ from it, however, in one of our research goals: we hope to bring the tools of computational morphophonemic learning to bear on a long-standing controversy in theoretical phonology. This is the so-called "abstractness controversy," which addresses the question of how far can URs diverge from surface representations (SR). Within phonology itself, 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 language learners. For our review of the abstractness controversy, see §2.

In earlier research on morphophonemics, authors have often assumed a particular position on the controversy — often either very concrete or fully 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 can be varied systematically. When we apply our learning system to language data, we keep the basic learning mechanism constant, while varying the principles that govern how abstract the URs are allowed to be. Our interest is in cases in which the very same learning system succeeds under one approach to abstractness, but fails at another, thus potentially shedding light on the abstractness controversy.

In pursuing this program, we have 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 the possible distance between underlying and

[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).

surface forms. We will refer to this series as the **KK Hierarchy** and make use of an important subset of it as the basis for evaluating the learnability implications of abstractness.[3]

The key idea is that several of the levels of the KK Hierarchy are directly translatable 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 (though always finite) set of candidate URs. To preview an example from §6.3 below, the Seediq stem for 'hold' appears on the surface in the allomorphs [pemux] and [pumex]. Given this set, our implementation of KK's "level D" generates the following set of possible URs for this stem: {/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 phonological constraint weightings, that successfully generates the Seediq paradigms. We express three other KK levels (KK-B″, KK-C, and KK-E) as UR candidate-creation algorithms as well.

Our work ultimately leads to a demonstration example (§6.6) of a type that we think will be informative in the long run: a case where a specific language ceases to be learnable when we move too high on the KK Hierarchy — the search space is sufficiently large and unconstrained that our system gets stuck in a local optimum. In this way, we show that, at least in principle, issues of computation can be made to bear on issues of abstractness.

The rest of the article is organized as follows. We begin with background sections on phonological abstractness (§2) and the KK Hierarchy (§3), as well as a toy data set (§4) used to illustrate our system. §5 covers our learning system, starting with morpheme parsing and culminating in the joint learning of URs and phonological grammar. §6 describes how the system deals with a series of examples at varying levels of the KK Hierarchy, culminating in an example in which going too high on the hierarchy blocks learning. In the final sections, we offer our conclusions (§7.1) along with future directions for research (§7.2).

2. The abstractness debate: where do things stand?

The era of phonological research centered on *SPE* (Chomsky and Halle 1968) was notable for the degree of abstractness it proposed to tolerate in underlying forms: *SPE* deployed pre-Vowel Shift URs for the English long vowels, as in /diviːn/ for [dɪˈvaɪn] *divine*; Foley (1965) set up abstract /eː, oː/ for the [e, o] of Spanish that fail to diphthongize under stress; Lightner (1965) proposed /ɪ, ʊ/ for vowels of Russian that alternate as [e/o ~ ∅]; and so on.

Not long after *SPE*, the "abstractness debate" began: scholars questioned whether highly abstract analyses actually match the internalized knowledge of native speakers; see for example Derwing (1973), Hooper (1976), and Tranel (1981). The work of Kiparsky (1968, 1973, 1982) was particularly influential; it outlined a number of possible limitations on abstractness backed by empirical arguments based on data from language change. However, not all scholars adopted

[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.

concretist views, and defenses of abstractness continued to appear (e.g. Gussmann 1980, Dresher 1981). A helpful review of the early research stages appears in Sommerstein (1977); for a more recent update see Baković et al. (2022).

Since the original controversy, new empirical work has been carried out, including work on the very languages held to support high degrees of abstractness. Such work suggests to us that an approach that rejects extreme abstractness is likely to be on the right track. We here follow up on two key 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: historical change never includes cases in which claimed abstract URs are "resurrected" as actual pronunciations. For example, there is no Slavic dialect in which Lightner's [ɪ], [ʊ] have been restored as distinct segments; no dialect of Hungarian where the [iː]'s that trigger exceptional back harmony have reverted to the putative underlying [ɯː] proposed 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 language 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" in the sense of Kiparsky (1982), apply to obscure an older phonological pattern, and abstract analyses generally recapitulate this history in synchronic terms. The key question 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 often referred to as *restructuring* (e.g., Bynon 1977). For example, in his original work, Kiparsky addressed the case of backness harmony, where in various languages the sound change ɯ > i created stems with surface [i] that take back harmony. In Mongolian (Kiparsky 1973:27-28) such stems have 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 falling between the years the language was originally studied by Newman (1944) and their own field study. Polish jer alternations have evolved from an abstractness-friendly system early on (recapitulated as synchrony by Gussmann 1980 and Rubach 1984) to a concretist system based on sonority sequencing (Gorecka 1988, Czaykowska-Higgins 1988, Jarosz 2008, Rysling 2016; for the similar Russian case see Gouskova 2012 and Rysling 2016). Brame's (1972) abstract-/ʕ/ analysis of Maltese phonology was subsequently thrown into doubt by the observations of Comrie (1986). Hoffman's (1973) analysis of Okpe with abstract /ɪ/ and /ʊ/ was thrown into doubt by evidence put forth by Archangeli and Pulleyblank (1994) that these sounds are actually realized in surface forms. For related Urhobo, Aziza (2008) sets forth a system with abstract /ɪ/, /ʊ/, and /ə/, then observes (p. 17) that for younger speakers it is breaking down. Even a relatively

modest form of abstractness — the "composite URs" discussed in §7.1.2 — likewise appear to be vulnerable to diachronic reanalysis; see discussion below.

Another change since the heyday of the abstractness debate has been an increase of interest in phonological learning, as studied with implemented computational models. This interest impinges on abstractness: abstract representations are not necessarily objectionable on their own terms, but need adequate justification, especially for how they might be acquired (Dresher 1981, Odden 2005:297, Baković 2009). Computational explorations of the conditions under which abstract URs are learnable can offer a more explicit evaluation of abstractness under this criterion. Indeed, the computational work 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; we find that under the system we propose there are benefits to using somewhat concrete representations. The key to our work is to run the same learning system under a variety of different conditions, each of them corresponding to a different degree of abstractness for URs being tolerated. This variety is provided for us, with full explicitness, by the KK Hierarchy.

3. The KK Hierarchy in outline

The very 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 that 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 the allophonic detail of nasalization on [æ]. We agree with Kenstowicz and Kisseberth (1977:9-11) that this is probably too restrictive. In what follows we will assume that UR learning makes use of representations that have already been phonemicized; e.g. our system as applied to English should "see" [pæn] and not [pæ̃n]. This enables the system to work with a much smaller segment inventory (the phoneme set) without loss of essential information. For textbook background on phonemicization see Hayes (2008:20-28), and for current computational models that address the phonemicization problem 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]

KK-B′′: The single-surface base hypothesis. In this approach, the language learner zeroes in on a particular slot in the paradigm (e.g., for Yiddish verbs, the 1st singular) and uses the allomorphs occupying that slot as the UR. This highly restrictive hypothesis has since been

[4] KK propose four additional levels, which for reasons of length are not be addressed here. KK-A′ says that only invariant features may be listed (alternating features are treated as unspecified). KK-B requires that the UR be based solely 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.

pursued by Albright and others, who demonstrate its capacity to explain historical change in Yiddish (Albright 2010) as well as Lakhota (Albright 2002), Latin (Albright 2005), Pengo (Dowd 2005), and Korean (Kang 2006, Albright 2008, Albright and Kang 2009). The hypothesis embodied by KK-B″ is so restrictive one might wonder why a learning system would even be useful; we give a reason in §7.1.4 below.

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. In other words, at KK-C, the set of UR candidates is identical to the phonemicized allomorph set.

KK-D: Segmentally-composite URs. Here, every segment in the UR must be realized faithfully in *some* surface form, but not necessarily the same form. Hence, URs can be "cobbled together" from more than one allomorph. An example is English /tɛlɛgɹæf/, cobbled together from *telegraph* [ˈtɛləˌgɹæf] and *telegraphy* [təˈlɛgɹəf-i] (*SPE*, pp. 11-12). Because the unstressed vowels in these words are neutralized to schwa, no single surface 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 such segments occurs in some allomorph. In the analysis of Yates (2017), Hittite phonology is an example of KK-E: he posits underlying vowels that are [+long, −accent]; these always surface either as [−long, −accent] or as [+long, +accent], depending on context, but never in their unchanged UR form.

As we consider these various levels, it is important to remember that the hypotheses they embody are not about languages, but about the language faculty: at most one of them can be true; and all of them need to be tested against multiple languages for their impact on learnability.

4. A toy example for illustration: Pseudo-German

We will illustrate the workings of 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: i.e., phonemic transcription, labels indicating what morphemes are present, but no morpheme boundaries per se. Stems are identified as such in the glosses. The glosses are given in the 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**]	$\text{turtle}_{\text{stem}}$	[lo**b**a]	$\text{turtle}_{\text{stem}}$ plur.
d.	[pap]	$\text{soup}_{\text{stem}}$	[papa]	$\text{soup}_{\text{stem}}$ plur.
e.	[mi**k**]	$\text{plane}_{\text{stem}}$	[mi**g**a]	$\text{plane}_{\text{stem}}$ plur.
f.	[bek]	$\text{beer}_{\text{stem}}$	[beka]	$\text{beer}_{\text{stem}}$ 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.

The phonology of Pseudo-German is familiar from many introductory phonology courses: alternating stems have final voiced obstruents underlyingly (/bed/, /lob/, /mig/), and their isolation forms are derived with a phonology that devoices obstruents in final position. 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

As noted in §1, the research literature on computational morphophonemic learning has become substantial, and it is clear that scholars have prioritized a variety of different goals for their modeling. We list below three important goals to which we aspire in this work.

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 require that a model be tested against more substantial data, with training sets 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 can increase our confidence that the model can ultimately be scaled up to match the experience of human children engaged in phonological acquisition.

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 1998, 2000; Jarosz 2006; Apoussidou 2006, 2007; Tesar 2014; Cotterell et al. 2015; Rasin and Katzir 2016). We offer the opinion that constraint-based frameworks such as OT are likely to have more explanatory value and utility than models based on rules or other schemata. The constraints of OT relate the content of language-specific grammars to broader principles grounded in typological and phonetic study. 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 (to appear).

Use of probability. Classical OT provides no basis for the analysis of gradient phenomena, such as free variation (single URs yielding multiple outputs; Labov 1969 et seq.) lexical frequency matching (mimicking of lexical frequency patterns in novel contexts; 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 and theories of possible constraints), but instead of ranking the constraints it assigns to every constraint a weight, reflecting its strength. On the basis of the weights and the pattern of constraint violations, a probability is computed for each candidate, using the formula in (11) 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 level of the KK Hierarchy has been chosen. The inputs to the system also include a set of hand-supplied constraints, to which the system assigns weights, thus forming a MaxEnt phonological grammar that, with suitable weights, can generate SRs from URs. The final task is to simultaneously choose the URs and assign phonological constraint weights. Success is defined by a combination of these choices that accurately generates the forms 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.[5] (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 (the strategy pursued by Jarosz 2006, Cotterell et al. 2015, Nelson 2019, Rasin et al. 2021).[6] Note that 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 1998, 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 widely-proposed 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. In other words, while 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. This means that bad parses can reveal themselves by exhibiting excess, unnecessary alternation. To give an example, suppose that for

[5] Many models attempt to detect morphemes in unlabeled training data; for overview see Hammarström and Borin (2010). We view such models as important for study of the earliest phases of acquisition, during infancy (§7.2.1). Our modeling, in contrast, is intended to represent the mental computations of older children, who are already able to detect morphemes and are addressing the more advanced task of learning how morphemes alternate phonologically.

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

Pseudo-German [beda] 'cat-plural', we consider the erroneous parse [be-da], 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 strongly dissimilar to the plural allomorphs that will arise in the paradigms of other stems in (1). In contrast, where the data are correctly parsed (e.g. 'cat' = [bet], 'plur.' = [a]) dissimilarity is limited to a minimum, in the case of 'cat' just the one-feature voicing difference between [d] and [t].

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 results of this procedure matches the linguist-preferred morpheme parse, and renders possible the later discovery of suitable underlying forms and phonological grammar. Since the morpheme parser is not the main focus of this article, we relegate a fuller description to the Appendix below.

We note lastly that discontinuous morphemes (from infixation, metathesis, or root-and-pattern morphology) are currently beyond the scope of our model; they involve a larger search space for morpheme parsing and also involve issues of how the morphology combines discontinous morphemes into words. For some proposals in this area see Wilson (2018) and Xu et al. (2020). "Blended" segments (from phonological coalescence) are likewise a frontier area, unaddressed here.

5.2 Allomorph dictionary and morpheme ordering

The next step, following work such as Tesar et al. (2003), Merchant (2008), and Tesar (2014), is to collect an allomorph set for every morpheme. In our system, these allomorphs are used to construct the candidate URs, in various 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 to each morpheme. Following this procedure for all morphemes, we obtain an **allomorph dictionary** for all forms. A subset of the allomorph dictionary for Pseudo-German is given in (2).

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

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

Moreover, by inspecting the ordering of the segments in the data, it is often straightforward to determine the principles of morpheme ordering.[7] For example, in Pseudo-German, all segments affiliated with a stem morpheme precede all segments affiliated with the plural morpheme, suggesting that plural is a suffix and must always follow the stem.

[7] In the cases we have studied 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; this suffices at least for the languages treated in this article, where morphemes are contiguous segment strings.

5.3 Detection of alternating segments

To learn the phonology, it will be necessary to discover what segments alternate. Here we make use of optimized string alignment. For instance, the 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).

(3) *Two representative optimal allomorph alignments*

a. Pseudo-German 'dog'			*b. Epenthesis/Syncope*			
b	e	d	a	p	t	ə
b	e	t	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 here 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 from the literature, presented, for instance, in Kruskal (1983) and in greater detail in the Appendix below. Previous uses of this method in phonology include 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])
[p] ~ [b]	(from [lop] ~ [lob])
[k] ~ [g]	(from [mik] ~ [mig])

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

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,[8] yielding (5) for Pseudo-German.

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

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

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

For higher levels on the KK Hierarchy, the choice of UR candidates is more complex, and we defer presentation until the particular examples of §6 below that illustrate these levels. It will emerge in most cases 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.[9]

5.4.1 Probability distributions over URs

The UR candidates are in competition with one another; usually, by the end of learning, one of them will dominate over the others.[10] Following earlier work (Jarosz 2006, Cotterell et al. 2015, O'Hara 2017) we represent this by assigning a probability distribution over all UR candidates affiliated with a particular morpheme, summing to one.[11] This distribution is altered over the course of learning. Table (6) gives partial results of our Pseudo-German learning simulation, listing the initial and final probabilities assigned to representative candidate URs.

(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.

5.4.2 Concatenated URs

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 μ_1, μ_2 … that comprise it; since all candidates are included, the probabilities of possible URs for each word will sum to 1.

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

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

[11] 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.

(7) *Whole-word UR candidates for Pseudo-German and their initial probability*

Word	*Candidate*	*Initial probability*
[beda]	/bed+a/	.5 × 1 = .5
	/bet+a/	.5 × 1 = .5
[bet]	/bed/	.5
	/bet/	.5

5.5 Phonological framework

For the reasons given in at the start of this section, the phonological portion of our system is expressed as a MaxEnt OT grammar (Goldwater and Johnson 2003), in which the intuitive "strength" of constraints is expressed using numerical weights rather than with ranking. For each language, appropriate constraints (mostly taken from the OT literature) are given to the learner in advance; for instance, in Pseudo-German, we hand-fed the model with the constraints used by Pater et al. (2012), namely *FINAL VOICED OBSTRUENT (abbreviation *d]), *INTERVOCALIC VOICELESS OBSTRUENT (abbreviation *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]:

(8) *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

In the tableau, Harmony ($\mathcal{H}$) is the weighted sum of all constraint violations; $e^{\mathcal{H}}$ is $exp(-\mathcal{H})$; i.e. e taken to the negative of Harmony; Z sums $e^{\mathcal{H}}$ over all candidates, and probability (p) for each candidate is its share in Z. The full equation for MaxEnt probability calculation appears in (11) below.

We adopt here a particular stance toward the question of 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, though related, component of the phonology as a whole; 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 (see 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.

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, as discussed in the following section, 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 et seq.) and stem-bounded phonotactics; see Paster for examples of the latter.

5.6 GEN for surface candidates

Our GEN function, taken in its essentials from Eisenstat (2009), is specifically suited to the learning of alternations. Eisenstat's method, which we will call **alternation-substitution**, takes advantage of the list of segmental alternations, which for us are obtained as in §5.3. Specifically, for each UR candidate, GEN is obtained by applying every change on the list wherever it is applicable, in all possible combinations. For instance, since [t] ~ [d] is in the alternation set for Pseudo-German, then in considering the UR /bed/, we must include the surface candidate [bet]. The presence of [t] ~ [d] in the alternation set also implies that 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. Many forms will include more than one location for substitution, and for these, all possible permutations are included in GEN.

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.

(9) *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, 3rd singulars, etc.: /fɪʃ-z/ → [fɪʃəz] 'fishes'. If we freely extended substitution of [ə] for zero, the candidate set would become infinite ([fɪʃəəz], [fɪʃəəəz], [əəəəəfɪʃəzəəəəə], etc.). Our treatment works as follows. The [ə] ~ ∅ alternation is discovered in the first place by the process of allomorph alignment described in §5.3 and illustrated below in (10):

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

ə	z	as in *fishes* [fɪʃəz]
∅	z	as in *dogs* [dɔgz]

The system observes that there is a [ə] ~ ∅ 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 create an additional candidate; thus, for the *possessive* form of 'fish', /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,[12] but it suffices for the examples considered.

We note that 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 phonological 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, sufficient to include the forms that permit repair for any input.

Our GEN system generally creates candidate sets that are smaller than those used in other systems (see, e.g. Ellison 1994, Eisner 1997, Albro 1998, 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 (i.e. that 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 phonological 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 (e.g. Tesar and Smolensky 1998, 2000; Tesar 2004; 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 solution that we adopt is as follows. First, we define an *objective function*, given below in (15), designed to reach its maximal value for the most accurate analysis. Second, to find the values of $\boldsymbol{\theta}$ and $\boldsymbol{W}$ that maximize the objective function, we employ the *Expectation-Maximization algorithm* (Dempster et al. 1977), an approach has been widely adopted in phonological learning (e.g. Jarosz 2006, Pater et al. 2012, Cotterell et al. 2015, Johnson et al. 2015, Nazarov and Pater 2017, Nelson 2019, Tan 2022, Pater and Prickett 2022).[13]

[12] 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.

[13] The particular deployment of key ideas we offer differs from all of these works; like Jarosz, Cotterell et al., and O'Hara, we employ probability distribution over candidate URs rather than affiliating candidate URs with UR constraints (cf. fn. 11); and like Pater et al. *et seq.*, we adopt MaxEnt as the phonological grammar, for the reasons given in §5. The combination of the two requires a novel deployment of EM in parameter estimation, described in this section.

In the remainder of this section we flesh out these ideas, giving the key formulae. Readers curious to see a worked-out example may consult the spreadsheet EMDemo_PseudoGerman.xlsx in the Supplementary Materials, in which formulae (11)-(18) are applied step by step to yield the correct URs and constraint weights for Pseudo-German.

Probability of an SR given a UR. This is computed by the MaxEnt phonological grammar, which consists of a set of constraints **C** (functions that assign violation counts to UR-SR pairs), along with their weights ***W***. Formula (11) (Goldwater and Johnson 2003:2) uses these constraints and weights to derive SR probabilities given a UR.

(11) *Computing the probability of a surface candidate given an underlying representation*

$$P(s \mid u;\ \boldsymbol{W}) = \frac{1}{\mathrm{Z}} exp(-\sum_i W_i C_i(u,s))$$

where $\mathrm{Z} = \sum_{s' \in \mathrm{GEN(u)}} exp(-\sum_i W_i C_i(u,s'))$

This formula computes the various expressions seen in tableau (8): Harmony ($\sum_i W_i C_i(u,s)$), $e\mathcal{H}$ ($exp(-\sum_i W_i C_i(u,s))$), and *Z*.

Aggregating the probability of surface forms with multiple URs. When multiple URs are in contention, the probability of a surface form *s* must be computed by forming a weighted sum of the probabilities of *s* being derived from any one of these URs; hence (12).

(12) *Probability of a surface candidate for a word*

$$\begin{aligned} P(s \mid \omega\ ; \boldsymbol{\theta}, \boldsymbol{W}) &= \sum_{u \in UR(\omega)} P(s,u \mid \omega;\ \boldsymbol{\theta}, \boldsymbol{W}) \\ &= \sum_{u \in UR(\omega)} P(s \mid u;\ \boldsymbol{W})\ P(u \mid \omega; \boldsymbol{\theta}) \\ &= \sum_{u \in UR(\omega)} P(s \mid u;\ \boldsymbol{W})\ \prod_{i=1}^{n} \theta_{(\mu_i, v_i)} \end{aligned}$$

where *u* is a possible UR for ω, consisting of the concatenation $(v_1 v_2 v_3 \ldots v_n)$ of the URs of its morphemes $(\mu_1 \mu_2 \mu_3 \ldots \mu_n)$.

The expansion of $P(u \mid \omega; \boldsymbol{\theta})$ in the third line expresses the idea (§5.4.1) that the probability of a candidate UR for a word is the product of the probabilities of the various URs for the morphemes that comprise it. Our use of the KK Hierarchy always yields a finite number of URs for each word, so that the summation in (12) is over a finite discrete set.

Calculating likelihood. The objective function is based on the conditional *likelihood* of the model as applied to the data. This is defined as the product of the probabilities assigned to every observation in *D*, under a particular setting of the model's parameters.

(13) *Calculating likelihood*

$$P(D \mid \boldsymbol{\theta}, \boldsymbol{W}) = \prod_{(s,\omega) \in D} P(s \mid \omega;\ \boldsymbol{\theta}, \boldsymbol{W})^{f(s,\omega)}$$

where $f(s, \omega)$ is the frequency with which *ω* is realized as *s* in the learning data

$P(s \mid \omega; \boldsymbol{\theta}, \boldsymbol{W})$ is the probability that the grammar assigns to s as the surface output for ω, as defined in (12).

Instead of maximizing likelihood directly, it is convenient instead to maximize the log of the likelihood, calculated as in (14).

(14) *Defining log likelihood*

$$ln(P(D \mid \boldsymbol{\theta}, \boldsymbol{W})) = \sum_{(s,\omega)\in D} f(s,\omega)\, ln\big(P(s \mid \omega;\ \boldsymbol{\theta}, \boldsymbol{W})\big)$$

Lastly, following standard practice, we augment (14) with a regularization term, which serves to avoid infinite weights and overfitting. With this term included, the objective function is now defined as in (15).

(15) *Defining the objective function*

$$L = \sum_{(s,\omega)\in D} f(s,\omega)\, ln\big(P(s \mid \omega;\ \boldsymbol{\theta}, \boldsymbol{W})\big) - \sum_i \frac{(w_i-\mu_i)^2}{2{\sigma_i}^2}$$

Optimizing θ and W with Expectation-Maximization. It is at this stage that the problem of hidden structure is addressed by using Expectation-Maximization. EM is an iterative method that breaks down an optimization task involving hidden structure into a set of smaller steps and alternates between them. The two steps are the Expectation (E) step and the Maximization (M) step.

E-step. The E-step fills in the missing UR information in the input by calculating expected values. Intuitively, it calculates how much "responsibility" each UR *u* should take for any observed datum (*s*, *ω*). This is done by calculating a posterior probability distribution over the hidden structures, using Bayes' Theorem, as shown in (16a).

(16) a. *The probability of the UR u, given observation* (*s*, *ω*)

$$P(u \mid s, \omega) = \frac{P(s|u,\omega)\, P(u|\ \omega)}{P(s \mid \omega)} \qquad \textit{by Bayes' Theorem}$$

$$= \frac{P(s|u)P(u|\ \omega)}{\sum_{u'\in UR(\omega)} P(s, u' \mid \omega)} \qquad \textit{by law of total probability}$$

$$= \frac{P(s|u)\, P(u|\ \omega)}{\sum_{u'\in UR(\omega)} P(s|u')P(u' \mid \omega)}$$

b. *E-step: expected frequencies of a UR in observation* (*s*, *ω*)

$$E(u, s, \omega) = f(s,\omega)\frac{P(s|u)\, P(u|\ \omega)}{\sum_{u'\in UR(\omega)} P(s|u')P(u' \mid \omega)}$$

where $P(s \mid u)$ and $P(s \mid u')$ are as given in (11).

M-steps. During the M-steps, the algorithm obtains a better estimate for the parameters $\boldsymbol{W}$ and $\boldsymbol{\theta}$ by maximizing the likelihood of "filled-in" data, based on the guess made at the Expectation step. The calculations for $\boldsymbol{W}$ are given in (17); those for $\boldsymbol{\theta}$ in (18).[14]

(17) *M-step: Using estimated frequencies to calculate constraint weights*

$$\boldsymbol{W}^{t+1} = argmax_{\boldsymbol{W}}\{\sum_{(s,\omega)\in D}\sum_{u\in UR(\omega)} E(u,s,\omega)\, ln\big(P(s \mid u;\boldsymbol{W})\big) - \sum_i \frac{(w_i-\mu_i)^2}{2{\sigma_i}^2}\}$$

where $P(s \mid u;\boldsymbol{W})$ is defined in (11)

(18) *M-step: Using estimated frequencies to re-estimate UR probabilities*

$$\theta_{(\mu,v)}^{t+1} = \frac{\sum_{(s,\omega)\in M(\mu)}\sum_{u\in UR(\omega)} E(u,s,\omega)}{\sum_{v'\in UR(\mu)}\sum_{(s,\omega)\in M(\mu)}\sum_{u'\in UR(\omega)} E(u',s,\omega)}$$

where μ is a morpheme, and v is a possible UR for μ
u, u' are members of the word form UR sets containing v or v'
M(μ) is the set of word forms containing μ,
E is as defined as in (16b).

The complete learning process. With both the E-step and the M-steps complete, a single iteration is accomplished. The next iteration begins by inputting the new parameter values $\boldsymbol{\theta}^{\boldsymbol{t+1}}$ and $\boldsymbol{W}^{\boldsymbol{t+1}}$ to (16b), and the process continues.[15] The overall procedure is this: the first iteration commences by assigning initial values $\boldsymbol{\theta}^{\boldsymbol{0}}$ and $\boldsymbol{W}^{\boldsymbol{0}}$ to $\boldsymbol{\theta}$ and $\boldsymbol{W}$, and the process terminates when an iteration ceases to improve log likelihood by more than some small threshold amount. We give the details for these settings in §6.

5.8 Results for Pseudo-German

The results obtained by our model for Pseudo-German are given in (19).

(19) *The course of learning for Pseudo-German*

a. *Weights* ($\boldsymbol{W}$)

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

[14] Formulae (17) and (18) are derived by differentiating (14); for discussion see Supplementary Materials.

[15] The discussion oversimplifies slightly, giving the classical layout of EM. Our own implementation achieves slightly faster convergence by carrying out the E-steps (16) not just before (17) but also before (18); see Meng and Rubin (1993), McLachlan and Krishnan (1997, ch. 5).

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
θ(‘plur.’, /a/)	1	1

c. *Sample final probabilities for surface candidates*

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

As can be seen, the learned values for ***θ*** and ***W*** suffice to select the linguist-expected URs and constraint weights with acceptable accuracy; i.e. the linguist’s answer for Pseudo-German is indeed the answer that is learned by the system.

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 the constraint weights assumed at that stage (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 is commonly noted as a disadvantage of this method (see e.g. Do and Batzoglou 2008, Cotterell et al. 2015, Tan 2022). However, in a research context, this tendency might actually be advantageous (cf. Nazarov and Pater 2017), since it can be made to bear on the choice of the empirically correct KK level. This line of inquiry is developed later in sections §6.6 and §7.1.

6. Case studies

We turn now to the evaluation of our model in a series of case studies. Our research design was to try to learn the same set of languages under a variety of KK levels, setting all other model parameters constant.

Here are the settings for our algorithm as we applied it in the language simulations. ***θ***: All candidate URs generated for a given morpheme (at a specific KK-level) were initially set as

equiprobable. ***W***: as is common in MaxEnt work (see e.g. Hayes and Wilson 2008, Nelson 2019, Pater and Prickett 2022), we started out all constraint weights with the value 1, and reimposed this default for each application of (17). For MaxEnt optimization there are multiple algorithms; we adopted the widely-used L-BFGS-B, described in Byrd et al. (1995) and Nazarov and Pater (2017). The values we employed for the regularization term in (17) were μ_i at 0 and $2\sigma_i^2$ at 10^5 for all constraints. Learning was terminated on the first iteration at which log-likelihood increased by less than 10^{-3}. Lastly, the phonological constraints we employ in our analyses are largely taken from McCarthy and Prince (1995).

Our first language example, drawn from Catalan (§6.1), is meant to show how our system permits learning at a scale corresponding to a typical first-year problem set; as it turns out, Catalan can be learned at any of the KK levels deployed here. We then present a series of smaller studies that require use of higher levels of the KK Hierarchy: Tangale, which requires at least KK-C (§6.2); Seediq, which requires at least KK-D (§6.3); and "Paka-20," which requires KK-E (§6.4). For the latter two cases, we provide a way that KK's ideas can be operationalized as algorithms generating candidate URs. In §6.6, we offer a new KK level of our own devising, and show that the abstract URs made available in this level can block successful learning in Catalan and Tangale.

For all of our examples (along with Pseudo-German above), we offer documentation of how the software implementing our algorithms did its work; this may be found in the Supplementary materials cited in the first footnote.

6.1 Catalan phonology

Catalan, a Romance language spoken in Catalonia and neighboring areas, is an ideal language for exploring UR learning at a scale characteristic of problem sets. Catalan phonology includes domains that involve lexical variation and opaque process interaction, offering additional challenges to learning systems. In addition, 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 that used by Cotterell et al. (2015), who in turn made use of a Catalan dataset from Kenstowicz and Kisseberth's (1979) textbook.[16] In order to include lexical variation, we added some additional paradigms from the sources cited above.[17] Our full set consists of 33 four-member paradigms in which stems are inflected for both gender (masculine -∅ and feminine [-ə][18]) and number (singular and plural), as exemplified in (20). We

[16] For other studies in learning Catalan alternations, see Shilen and Wilson (2022) and Ellis et al. (2022).

[17] 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.

[18] 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. We opt here for a

include morpheme boundaries for clarity, though they are not present in the data presented to our system.

(20) *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’
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,[19] whereas in the masculine forms various types of phonology attack the exposed coda consonants.

Final cluster simplification. This can be seen in paradigms such as (20e) and (20f) above, 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), banning skipping in surface realizations.

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 (20g) and (20h). As Wheeler (2005:327-338) emphasizes, the loss of singleton consonants is not a regular, across-the-board process, but involves many exceptions. We have included in our training set both deleting and non-deleting

morphological analysis in which the feminine vowel is underlyingly uniform and the plural is [-s] added to the singular, as in the masculine. As far as we can tell, nothing crucial hinges on this choice.

[19] Feminines do involve Spirantization of the stem-final consonant, as in [kazað-ə-s] ‘married f.sg.’, stem UR /kazad/. Following the assumption given above (§3), the input to learning consists 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/.

stems for both /n/ and /r/, selecting (as an approximation) a ratio of three deleting stems for each non-deleting stem.[20]

(21) *Dropping of singleton final consonants — forms that undergo, and non-undergoing forms*

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'

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**	pur-s	pur-ə	pur-ə-s	'pure'

We agree with Wheeler (p. 330) and much subsequent work that unpredictable forms in cases like this are 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 acquisition, adults encountering novel stems, or a wug test. Wug-testing in other languages suggests that, all else being equal, adult speakers are likely to *frequency-match* such patterns (Zuraw 2000, Ernestus and Baayen 2003, Albright and Hayes 2003, Pierrehumbert 2006, Hayes et al. 2009, Becker 2009, Gouskova and Becker 2013).[21] Hence if our output grammar is able to achieve 75% probability of deletion for both /n/ and /r/, we will regard it correct for present purposes.[22]

The processes of singleton deletion in (21) are not just exceptionful, but also opaque: as expressed in rule-based phonology /n/ Deletion and /r/ Deletion would be *counterfed* by Cluster Simplification. The grammar must guarantee this counterfeeding effect, since there are no alternations at all in Catalan like *[sant-ə] ~ [sa]. For this purpose, we adopt a highly weighted Faithfulness constraint banning alternations of the form *CC ~ ∅, roughly following Kirchner (1996).

[20] The lexical survey for Liang et al. (in preparation) suggest 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.

[21] 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).

[22] To our knowledge, no published wug test study of Catalan has appeared. However, in an ongoing experimental study, Liang et al. (in preparation) find that Catalan wug-testees given hypothetical stems ending in /n/ or /r/ do indeed delete the consonant some of the time, but not all of the time (79.7% deletion for /n/, 52.4% for /r/). Participants also respect particular statistical tendencies observed in Mascaró (1976) and Wheeler (2005): more deletion in listed morphemes like /-dor/ 'agentive', less deletion in monosyllables and far less deletion in paroxytonic stems.

Epenthesis. The data in (20k)-(20n) 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 also has a devoicing process similar to Pseudo-German, illustrated in (20j)-(20m); for example /griz/ → [gris]. Cases like (20k) /seg-s/ → [sek-s] suggest that devoicing applies to all coda obstruents, not just word-final ones, so we use the constraint *CODA VOICED OBSTRUENT. This constraint 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.

There is, however, an interesting additional wrinkle, in that final /ʒ/ surfaces in final position not as the expected [ʃ] but as [tʃ], as in /bɔʒ/ → [bɔ**tʃ**], from (20l). This is a case of "saltation" (Lubowicz 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 to saltation given in Hayes and White (2015), banning the expected-but-unwanted "short" journey using a *MAP constraint (Zuraw 2007, 2013), which bans a particular segmental correspondence (going in any direction). Here, *MAP(ʒ ~ ʃ), weighted higher than *MAP(ʒ ~ tʃ), forces the victory of [tʃ] in [bɔtʒ]. According to 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 does indeed tend to lead to repair by Catalan language learners. In the analysis, we included several *MAP constraints definable over the four segments [ʃ, ʒ, tʃ, dʒ].

In sum, our account of this subset of Catalan phonology makes use of the constraints in (22), which are listed with the weights that are ultimately assigned to them by our learning system.

(22) *Phonological 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 ~ ∅	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 assuming the UR-candidate selection procedure of KK-C (all and only allomorphs are candidates), but in this particular case the choice of level does not matter, unlike in the cases below: all four KK levels of §3 suffice to find the right analysis.[23]

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]; and 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. Allomorph alignment (§5.3) also correctly located all of the segmental alternations in the full dataset, namely { 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 the alternation-substitution method (§5.6) to create surface GEN, yielding a total of 14,130 UR-SR pairs. The hand-provided constraints were as given in (22). 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 prevocalic (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 given in (22) above. (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 very close match to the patterns present in the training set.

We note further that the learner attributed *all* the variation to the grammar, rather than to multiple URs. For example, while 'good' surfaces as both [bo] and [bon], the learner assigned

[23] 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.

essentially 100% probability to /bon/, letting [bo] be derived solely 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 set at every level of description.

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 of our model easily provides the correct parse when given the paradigm of a novel stem; it does this by searching the parses for just that stem, with the parses for known vocabulary fixed in place. For example, [primerəs], not in the original training set, is correctly parsed as [primer-ə-s] 'first-fem.-plur.' Moreover, given the parsed paradigm for this stem ([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 of all five languages discussed here; a sample demonstration for /primer/ is given in the Supplementary Materials.

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 the textbook by Kenstowicz (1994), studied before by Cotterell et al (2015).[24] The data consist of nouns occurring both alone and with five different suffixes.

(23) *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'

[24] Following Cotterell et al. (2015), we omitted tone and the phonemic ATR distinction.

These data exemplify four phonological processes, analyzed in rule-based phonology by Kidda and Kenstowicz: (a) **Final Devoicing** of obstruents; thus (23c) /tugad/ → [tugat] 'berry'; cf. [tugad-i]. The stem in (23b), [bugat] ~ [bugat-i] illustrates a stem ending in underlyingly /t/. (b) **Syncope**: stem-final short vowels are deleted before a suffix, as in (23f) /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 (23k) /lutu-go/ → lutgo → [lutko]. (d) **Pre-sonorant voicing**: Obstruents placed before 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 specifically requires us to adopt 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 (23j) /ŋuli/, only the isolation form can tell us 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 (24).

(24) *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 in the style of Kirchner (1996); forces opacity (no 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[25]

[25] The value of 4.36 seen here is surprisingly high, but harmlessly so, since the countervailing constraints IDENT(voice)-stem and IDENT(voice) combine for a far higher Harmony value (12.59), preventing intervocalic voicing. *VTV is in fact not 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] and /loː-go/ → *[loː-ko].

REALIZEMORPH	Violated when a morpheme is not realized by any segment in the surface form; prevents /duka-i/ → *[duka] for (23l)	10.43

With the learning data of (23), the constraints of (24), and the assumption of KK-C, learning proceeded without incident, yielding the results shown in (25).

(25) *Tangale learning results*

a. *URs:* all correct URs assigned probability > 0.999
b. *Constraint weights:* as shown in (24)
c. *Surface forms:* All correct forms assigned probability > 0.999

Most crucially, the system properly seized upon the isolation form (e.g. /ŋuli/) as the basis for vowel stems, but the contextual form (e.g. /tugad-i/) for obstruent final stems, rather than consistently sticking with any particular paradigm slot, as KK-B″ would require. This makes it possible to derive the correct outcomes. We return to Tangale below (§6.6), where we use it to show that ascending too high on the KK Hierarchy can also lead to learning failure.

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

Seediq is an Austronesian language of Taiwan, described and analyzed in Kuo (2020, 2023), who builds on earlier work by Yang (1976). Seediq provides a canonical illustration of 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. Crucially, a neutralizing process (vowel reduction) applies to every paradigm member of a stem, but not in the same location. Therefore, in classical analysis the UR must be "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; later on (§7.1.2), we explore the possibility that the classical analysis is not necessarily the correct one.

Some representative data are given in (26).

(26) *Vowel reduction in Seediq*

Vowels	*Composite UR*	*no suffix*	*1 suffix*[26]	*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**ur**ˈh-an-i]	'rare'
c. /ai/	/galiq/	[ˈgaliq]	[g**u**ˈliq-an]	[g**ul**ˈq-an-i]	'break'
d. /ea/	/gedaŋ/	[ˈgedaŋ]	[g**u**ˈdaŋ-an]	[g**ud**ˈŋ-an-i]	'die'

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

e. /ua/	/burah/	[ˈburah]	[b**u**ˈrah-an]	[b**u**r**u**ˈh-an-i]	‘new, create’

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. The process of vowel reduction works as follows. If a vowel of any quality is *pretonic*, it is realized as [u], as in (26a-e). If a vowel is *posttonic and mid* ([e] or [o]), it is realized as [u], as in (26f-g). Since stress migrates within the paradigm, based on the number of suffixes present, different vowels of a stem will be targeted by reduction depending on the paradigm slot. The specific need for composite URs is demonstrated by forms like (26f), /pemex/, where we need to consult the isolation allomorph to know that the UR has /e/, not /u/, as its first vowel; and we need to consult the single-suffix allomorph to know that the UR has /e/, not /u/ as its second vowel. This style of analysis has often been used for similar languages in which stress is mobile within paradigms and stressless vowels are reduced or deleted (English: Chomsky and Halle (1968:11-12), Palauan: Flora (1974), Tonkawa: Kisseberth (1970), 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.[27] Our training data consists of 15 stems (45 words), covering enough vowel patterns to test out the constraint system.

A simple set of constraints suffices for the analysis. The cover constraint PENULT STRESS enforces stress on penults and stresslessness elsewhere; it stands in for a set of appropriate constraints demanding a single word-final trochaic foot. The Markedness constraint PRETONIC REDUCTION bans all vowels but [u] in pretonic position, and POSTTONIC REDUCTION analogously 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] ([i] is legal posttonically, as seen in (26c)). Hence we adopt the counternatural weighting of *MAP(e, i) over *MAP(e, u); we also include a relatively complete set of other *MAP constraints, though the analysis would work without them.

(27) *Phonological constraints for Seediq*

Constraint	*Characterization*	*Weight in learned analysis*
PENULT STRESS	Cover constraint, standing 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

[27] 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, since stress is not phonemic;[28] included for concreteness.	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[29] 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 (26f-g), no one allomorph encodes all of the underlying vowels. The learner cannot find the classical solution at KK-C 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 level 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, (28a) gives the calculated optimal alignment for two surface allomorphs of (26f-g) ‘hold’. The needed set of URs can be obtained simply 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 (28b).

(28) *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

[28] 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.

[29] *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 (28) 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 (29).

(29) *Seediq: final learning results*

a. *URs:* all correct URs assigned probability > 0.999
b. *Constraint weights:* as in (27) 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 intended to facilitate study of phonological learning algorithms, including algorithms that learn URs. Since then, Paka languages have become a focus of formal work in phonological learnability; see e.g. Alderete et al. (2005), Jarosz (2006), Merchant (2008), Tesar (2014), DelBusso (2020), and Tan (2022). In light of this body of work, we see the ability to cover Paka languages as an essential criterion for evaluating proposals in UR learning, and therefore checked to make sure that our system can handle all 24 of the languages (i.e, in the version of Paka given in 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 (30).[30]

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

(30) *Paka-20 paradigms*

UR	*SR*	*Gloss*	*UR*	*SR*	*Gloss*
/pa-pa/	[pápa]	‘dog-nom.’	/ʧé-pa/	[ʧépa]	‘pig-nom.’
/pa-tiː/	[páti]	‘dog-acc.’	/ʧé-tiː/	[ʧéti]	‘pig-acc.’
/pa-ʧé/	[paʧé]	‘dog-gen.’	/ʧé-ʧé/	[ʧéʧe]	‘pig-gen.’
/pa-kóː/	[pakóː]	‘dog-abl.’	/ʧé-kóː/	[ʧéko]	‘pig-abl.’
/tiː-pa/	[tíːpa]	‘cat-nom.’	/kóː-pa/	[kóːpa]	‘bat-nom.’
/tiː-tiː/	[tíːti]	‘cat-acc.’	/kóː-tiː/	[kóːti]	‘bat-acc.’
/tiː-ʧé/	[tiʧé]	‘cat-gen.’	/kóː-ʧé/	[kóːʧe]	‘bat-gen.’
/tiː-kóː/	[tikóː]	‘cat-abl.’	/kóː-kóː/	[kóːko]	‘bat-abl.’

In Paka-20, morphemes contrast for underlying stress. 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 for ‘dog’ and ‘cat’ appear with surface accent only in the nominative and accusative, where the suffix is unaccented; whereas the underlyingly accented stems for ‘pig’ and ‘bat’ are invariantly accented. Moreover, vowel length is phonemic, as shown by the differences between ‘dog’ and ‘cat’, ‘pig’ and ‘bat’, and genitive and ablative. Underlying long vowels are shortened if they fail to receive stress, as we see for the stem vowel in the genitive and ablative forms of ‘cat’, and for the ablative suffix when it follows an accented stem.[31]

Is there a real language that patterns like Paka-20? The evidence of which we are aware suggest this is possible but not certain. In the analysis of Yates (2017), Cupeño (Takic, California) is an instantiation of 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 that is needed to detect phonology in the Hittite orthography makes the analytic conclusions less certain.

For the analysis of Paka-20, we follow in broad outlines the treatment in Tesar (2014:176). We found that, with the more complete set of URs and surface candidates generated by our system, we needed to specify the constraint set more fully. However, the alterations are minor and do not affect the point at hand.[32] Our constraints are given in (31), along with the weights that were obtained in our learning simulation.

[31] The reader may have noticed that the paradigms of (30) give no evidence that accusative /-tiː/ actually has an underlying long vowel; it is in effect a member of the Rich Base 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.

[32] Specifically, MAINLEFT should be defined as ALIGN(σ́, L, Word, L) “The left edge of every stressed syllable must coincide with the left edge of a word,” following the Alignment system of McCarthy and Prince (1993). We also added the standard constraints *CLASH (in order to exclude candidates that have two stresses) and *LAPSE (in order to exclude stressless candidates).

(31) *Phonological constraints for Paka-20*

Constraint	*Characterization*	*Weight in learned analysis*
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 using classical OT. Specifically, MAINLEFT is weighted much higher than MAINRIGHT (default choice for stress); IDENT(stress) bears enough weight to overcome MAINLEFT (so that in /pa-ʧé/, [paʧé] defeats doubly-unfaithful *[páʧe]); 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 will survive.)

We now explain why Paka-20 necessitates a still higher level of the KK hierarchy. The key paradigm is for 'cat', which requires us to set up the UR /tiː/ for the stem. This UR contains a segment, long unaccented /iː/, which never occurs in surface representations. The evidence for the existence of /iː/ in the UR comes from the phonological *behavior* of this stem. 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 contrasting 'bat'. Assuming URs like /tiː/, Paka-20 will defeat all versions of the KK Hierarchy thus far discussed, since all of these versions require that a segment occur in surface representation for it to be a possible UR segment. Paka-20 is a canonical instance requiring us to ascend to KK-E (§3), which allows URs to contain "featurally-composite" segments.

We think a sensible way of finding and including /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 in economical fashion because it is phonetically intermediate between vowels found in overt UR candidates, 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 (32), based on the similar definitions given in Tesar (2014) and Magri (2018:580):

(32) *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*.[33]

[33] 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

With this definition, we can provide a criterion of UR inclusion that can be used to form UR candidates at level E of the KK Hierarchy; this is given in (33).

(33) *Interpolation (KK-E)*

Let *x* and *y* be two aligned segments of two UR candidates, occupying the same column C in their aligned form, as in (28b). If *z* is a segment 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*.

In the present case, interpolation adds to the candidate UR set for 'cat' the two additional UR candidates /tiː/ and /tí/, both of which are intermediate between /tíː/ and /ti/. 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.[34]

We omit the full narrative of how our model worked here, since it is so similar to previous cases, but simply observe that it obtained the correct weights, UR probabilities (including ~1 for /tiː/), and correct probability (~1) for observed outputs). The details may be viewed in the Supplementary Materials. The key point is that cases like Paka-20 — if it is true that such cases exist and are psychologically real to their speakers — can be treated by a fairly straightforward expansion of the UR candidate set, thus 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.

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 embark on discussion of the implications of our KK-based system for abstractness in phonology.

We first wish to emphasize that the system we have set up 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 principled reason why our system should not be able to find a successful grammar.

---

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, 2018) for relevant discussion.

[34] 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. Saltation repair has actually been noticed in Catalan: in the wug-test study of Liang et al. (in preparation), participants often inflected wug stems such as [ʎuˈðaʒ-ə] 'wug-f.' with [ʃ], as in [ʎuˈðaʃ] 'wug-m.', rather than with the saltated [ʎuˈðatʃ] that would be expected from the Catalan lexicon.

We have checked this possibility in the case of O'Hara's (2017) abstract analysis of Klamath phonology. O'Hara proposes an abstract /e/ to underlie certain instances of [i] that alternate with ∅; /e/ contrasts with /i/ (which underlies cases of [i] that do not so alternate) and also with ∅. Abstract /e/ never surfaces, though there is an /e/ phoneme that is attested in other environments. None of the methods for UR hypothesis formation given above would ever 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 possible 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 surface data; for details see Supplementary Materials.

However, hand-including URs in the search space skirts what we take to be 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 to this issue in the following section.

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

For illustrative purposes, we invented our own new KK level, intended to test the effects of using a larger hypothesis space; we will call this level "KK-Z." To form KK-Z, we collect all allomorphs (as in KK-C), then apply the method of alternation-substitution. This method was used in §5.6 to create surface GEN candidates, but for KK-Z we apply it to the creation of UR candidates as well. For instance, because Tangale has both voicing and vowel ~ ∅ alternations, the allomorph set for 'berry' {[tugad], [tugat]} gives rise to $2^5 = 32$ possible URs; 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.[35] We employ it here solely to show that the restrictiveness of the hypothesis space can matter in the present context. Consider in particular what happens when we return to our Tangale example (§6.2), this time using KK-Z. We have found that the larger hypothesis space created under KK-Z led to serious learning failures, as follows: (a) aberrant voicing alternations: what by the standard analysis is /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: what by the standard analysis is /tugad/ surfaces as 25% correct surface [tugat] and 75% incorrect *[tugado], with epenthesis.

[35] 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 that we might call KK-EZ, which appears to be capable of locating the necessary URs for a number of well-known highly abstract analyses. For instance, KK-EZ could discover /beːn/ for English *bean* [biːn] (*SPE* p. 69), since /eː/ is plausibly 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 already suffices.

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 that DEP should be strengthened, 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.[36]

The situation with Catalan is similar: the grammar mishandles the [r] ~ ∅ alternation, treating cases like (20h) [kla] ~ [klar-ə] as stem allomorphy (the URs /klar/ and /kla/ are equiprobable). At the same time, the phonological grammar learns no process of [r] ~ ∅ alternation at all, with zero weight for *CODA [r] and high weights for MAX and DEP. The fatal 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 described in §6. However, we obtain the same outcomes under a different plausible initialization (cf. Smolensky 1996) in which Markedness constraints start out high (50) and Faithfulness low (1).

It is worth trying to diagnose these learning failures in general terms. We conjecture that when the learning system is confronted with a very large array of URs, as with KK-Z, it is liable to embark on the (false) task of making sure that the more remote URs map onto the 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 conservative KK levels, would never have arisen in the first place. Once the system has settled at a local optimum, it is defeated, frozen in place by the need to solve pseudo-problems.[37]

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 — involving pernicious "bumpiness" in log likelihood — that makes it too difficult to search in the cases of Tangale and Catalan. We turn to our interpretation of this result in the following sections.

[36] Note that if we hand-enter the correct $\boldsymbol{\theta}$ values for the UR candidates created under KK-Z, our system discovers constraint weights that derive the right surface forms. Moreover, if we hand-enter the constraint weights $\boldsymbol{W}$ for the correct grammar, the system discovers appropriate $\boldsymbol{\theta}$ values, again deriving the correct outcomes. It is the mutual dependence of $\boldsymbol{\theta}$ and $\boldsymbol{W}$ that leads the system astray.

[37] As our reviewers pointed out, perhaps some of the difficult cases might be fixed by invoking further principles of phonological theory. In particular, if for Tangale, Universal Grammar always 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 here, namely *[tugatko] (Tangale) and [kla]/*[klar] (Catalan). 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. Discussion

In discussing our results we first take on issues specifically related to abstractness, then cover further issues.

7.1 Relation to abstractness

To review what has been presented: we have shown that our proposed learning system, when provided with a phonological constraint set, can learn morphophonemic systems of the scope and complexity level seen in ordinary problem sets. The key ingredients are (a) early detection of morpheme membership by similarity, so that the allomorph sets can be found; (b) use of alignment to locate the alternating segments; (c) algorithms that project UR candidate sets at each level of the KK Hierarchy; and (d) use of Expectation-Maximization to find 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 pilot 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 — i.e., carries out restructuring — where young humans do the same. The UR abstractness controversy is founded on this point.

7.1.2 What is the right KK-level?

Under the criterion just given, the appropriate KK level for human learning should be empirically diagnosed by the occurrence of restructuring when sound change dishes up to children a pattern that exceeds the capacity of their learning system. This diagnostic can be applied in multiple areas, including opacity (Kiparsky 1971) and structural complexity (Moreton and Pater 2012). For the specific case of UR abstractness, we can 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 that is most likely to be correct, our guess would be KK-C (UR candidate set identified with allomorph set) or lower. The reason this is plausible is that in recent years, evidence has been gathered suggesting 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. Literature making this point includes work on English (Hayes 1995); Yidiŋ (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 early work done in the post-*SPE* era by 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, which has resulted in diachronic changes for a number of stems. As Kuo shows, these are the stems (like (26f-g) above) that would require KK-D-type URs under the analysis put forth above. The restructuring worked as follows: the underlying form is now identical to the isolation form, and the "restored" vowel that appears when stress is shifted does not come from the 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). Evidently, the vowel-copying generalization was already present on a statistical basis in Seediq even before vowel reduction entered the language; the later restructuring consisted of extending the principle of 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, but this time including Kuo's constraint (2023:5) NUC-IDENT-OO[F], which is responsible for the vowel copying effect. We also use a more realistic training set than before, with sufficient stems to represent the preference for vowel copying among existing forms.[38] We also added in a small number of hypothetical forms, meant to approximate stem types like / u e / that originally existed in Seediq but have been ironed out by restructuring. By this means we expanded the training set 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", with very close matching of 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 the Supplementary Materials.

[38] We would like to thank Jennifer Kuo for providing us with a dataset of 344 paradigms.

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.[39]

7.1.4 Informativeness and KK-B″

We have said little so far about the most concrete level of all, namely KK-B″, the single surface-base hypothesis. Albright's proposal (see references in §3) is that in early morphophonemic learning, children try out a variety of paradigm slots to serve as the basis for the UR, then ultimately settle on the most effective one and use it exclusively thereafter. We suggest that this early selection process might effectively be modeled at the level of KK-C, where multiple stem allomorphs compete to serve as the optimal 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 language learner would subsequently rely on this choice for all 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 in our system 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 (20k), 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.

[39] 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.

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). The human learner must also apprehend what morphological categories are present in the ambient data. For some first efforts to address these demanding tasks, see Baroni et al. (2002), Dreyer and Eisner (2011), Jin et al. (2020) and Wiemerslage et al. (2021).

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, Calderone et al. 2021, Wilson and Li 2021), but such tests have often employed models that do not make use of a level of underlying representation.

The reason our system would not do well on a wug test is that, like other UR-based models, it assumes optimum learning conditions; i.e. for any given morpheme, it is presented with a complete paradigm and asked to infer the best 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 target language, the isolation form is not the optimal form for guessing URs, since the language has Final Devoicing, but Dutch speakers can guess the UR of an isolation form at 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.[40]

7.2.3 Allomorphy, irregularity, inflectional classes

Our system can solve phonology problem sets, which are, perhaps, notorious for being cleansed of all patterns not attributable to productive phonology. These patterns 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).

[40] We speculate that fine-grained paradigm-guessing results like Ernestus and Baayen (2003) reflect the capacities of adult speakers, who have had the time to engage in paradigm learning in far greater detail than at the level described here.

By this view, an adequate future theory of morphophonemic learning must embed the more purely phonological part into a larger theory that handles the many aspects of morphology that are not phonologically reducible.

7.2.4 Learning the constraints

A more ambitious version of our learner would provide its own constraints, rather than relying on hand-provided ones. On the Markedness side, the obvious source for learned 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, Jardine and McMullin 2017, 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 the alternation-substitution form of 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 more ambitious "freedom of analysis" approach of standard OT (see, e.g., McCarthy 2007), which aspires also to cover phonotactics. Under the 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 conjoined constraint IDENT(voice) & MAX(V) that we used to account 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 due to derivations like /sant/ → [san]; these would likewise need additional mechanisms, or grammar architectures (Nazarov and Pater 2017), to be learned.

7.2.5 More powerful learners?

We are curious whether or not a computational system could be devised that, unlike ours, could learn effectively at high, abstractness-friendly KK levels. For instance, we noted above (fn. 35) 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.

Appendix: Description of the morpheme parser

To review, we assume a set of data like in (1), where words are segmental strings annotated with an indication of what morphemes are present. The task at hand is to find, prior to phonological discovery, which morpheme each segment belongs to. We offer one way to do this, which works for the data sets studied in this article. To our knowledge this is the first effort to achieve this goal, and it is likely that many alternatives might fruitfully be explored.

Our system is based on enumerating all the logically possible morpheme parses and choosing the best one as defined in a particular way.[41] The enumeration of candidate parses works as follows. If word *w* in the training data contains *n* morphemes, then for each segment ~~*s*~~ in *w* there will be *n* possible morphemic affiliations. If word *w* contains *m* segments and *n* morphemes, then the number of possible morphemic affiliations for *w* will be m^n. Then, for the full set of words w_i in the training data, we cross-classify all of these morphemic affiliations, obtaining $\prod_i m_i^{n_i}$ candidates.

As already discussed in §5.1, the best morpheme parse is likely to be the one that maximizes paradigm uniformity; i.e. minimizes the degree of divergence between the allomorphs of the morphemes present in the data. Thus, to evaluate any particular candidate, we assign it a penalty score that reflects this divergence. This score is built up piecewise from the segment level to the whole-parse level.

To start, we assess the dissimilarity of segment pairs. The difference is computed on the basis of their mismatching feature values, using a fairly standard phonological feature set. Each feature contributes a specific degree of dissimilarity penalty (for example, mismatches in [voice] contribute 3.4). The feature-specific penalties were established using the method of Wilson (2006) and White (2017): they employed the values that provide the best fit to confusion matrix data gathered experimentally (we used the data in Cutler et al. 2004). For any two segments, the dissimilarity is calculated as the sum of the penalties for every feature in which they differ.

The next task is to compute allomorph dissimilarity. We first compute for any given pair of allomorphs the optimal segment-to-segment alignment for that pair, using the method from Kruskal (1983) already described in §5.3. The method requires penalty weights for both (i) correspondence of non-identical segments, and (ii) corrrespondence of a segment to null. For (i), we used the values of segment dissimilarity just described, and for (ii) we adopted an *ad hoc* value, namely 45. A sample optimized alignment is shown in (34a) for the Pseudo-German allomorphs [bet] and [bed]. Here, the dissimilarity penalty comes solely from the difference of [t] and [d], 3.4. The dissimilarity of [apt] and [aptə], shown in (34b), derives from the segment-to-null penalty, 45.

[41] Our procedure is strongly reminiscent of classical (non-stochastic) Harmonic Grammar (Legendre et al. 1990), with the constraints drawn from the Correspondence Theory of McCarthy and Prince (1995). However, since our system is a procedure for morpheme discovery, not a phonological grammar, we present it here in neutral terms to avoid confusion.

(34) *Three representative best allomorph alignments with penalties*

a. Pseudo-German 'dog'

b	e	d	
b	e	t	
0	0	3.4	penalty weights
	3.4		total penalty

b. Epenthesis/Syncope

a	p	t	ə		
a	p	t	Ø		
0	0	0	45	penalty weights	
		45		total penalty	

It should be clear that any other alignment for these cases would incur a much higher penalty; for instance the nonoptimal alignment of (34a) discussed in §5.3, [b]-Ø, [e]-[b], [d]-[e], [Ø]-[t], with superfluous correspondence to null, would incur a very large penalty (124.9).

Finding the winning parse. Starting with the dissimilarity values for each possible pair of distinct allomorphs, we scale up as follows. The dissimilarity penalty for a *morpheme* is the average of the penalties for all the allomorph pairs in which it appears, rescaled to penalize alternation in stems more severely. The dissimilarity penalty for the *entire parse* (i.e. of the whole dataset) is the sum of the penalties for the set of morphemes it contains (no frequency weighting). The winning parse is the one with the lowest overall penalty. We note that the number of parses can in practice be very large (for Catalan, about 2×10^{161}). Such sets cannot be searched exhaustively, so instead we employ a greedy hill-climbing search in which candidates are edited in various ways, not covered here. This system suffices to obtain the intuitively correct morpheme parse for all of the language simulations mentioned in this article.

Directions for improvement. We noted above that our system attempts to solve a problem not previously addressed in the computational literature, and we judge that further attempts to sole it would be beneficial. Improvements might be possible on at least three lines. (a) The metric of similarity for allomorphs might be more accurate if it considered the *contexts* in which segments occur, since this seems to be right for similar problems in phonology (Steriade 2001, Fleischhacker 2005). (b) It seems likely that our search for the best parse could be carried out more efficiently. (c) The system could be made more flexible by having it forward to phonological learning a *set* of likely parses rather than just one. These could make the system tolerant of error at a modest cost in total search space.

References

Albright, Adam. 2002. A restricted model of UR discovery: Evidence from Lakhota. Paper given at the GLOW Phonology Workshop, Utrecht. http://web.mit.edu/albright/www/papers/Albright-Lakhota.pdf

Albright, Adam. 2005. The morphological basis of paradigm leveling. In *Paradigms in phonological theory*, eds. 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. Jeff Good, 144–181. London: Oxford University Press.

Albright, Adam. 2010. Base-driven leveling in Yiddish verb paradigms. *Natural Language & 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.

Albright, Adam, and Yoonjung Kang. 2009. Predicting innovative alternations in Korean verb paradigms. *Current issues in unity and diversity of languages: Collection of the papers selected from the CIL 18, held at Korea University in Seoul*, 1-20.

Albro, Daniel M. 1998. Evaluation, implementation, and extension of primitive Optimality Theory. Master's thesis, UCLA, Los Angeles, CA.

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

Alderete, John, Adrian Brasoveanu, Nazarré Merchant, Alan Prince, and Bruce Tesar. 2005. Contrast analysis aids the learning of phonological underlying forms. In *Proceedings of WCCFL*, vol. 24, pp. 34-42.

Apoussidou, Diana. 2006. On-line learning of underlying forms. Rutgers Optimality Archive ROA-835.

Apoussidou, Diana. 2007. *The learnability of metrical phonology*. Ph.D. thesis, University of Amsterdam.

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

Baayen, Harald, Robert Schreuder, Nivja de Jong, and Andrea Krott. 2002. Dutch inflection: the rules that prove the exception. In Sieb Nooteboom, S. G. Nooteboom, Fred Weerman, and F. N. K. Wijnen, eds., *Storage and computation in the language faculty*, 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 Lila Gleitman, Anna Papafragou, and John Trueswell, editors, *Oxford Handbook of the Mental Lexicon*. 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*, pages 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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6176–6186, Hong Kong, China. Association for Computational Linguistics.

Becker, Michael. 2009. Phonological Trends In The Lexicon: The Role Of Constraints. Ph.D. dissertation, University of Massachusetts Amherst. Amherst, MA.

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. Ph.D. dissertation, UCLA, Los Angeles, CA.
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 Michael Brame, ed., *Contributions to generative phonology*, 22-61. Austin: University of Texas Press.
Bynon, Theodora. 1977. *Historical Linguistics*. Cambridge: Cambridge University Press.
Byrd, Richard H., Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16. 1190–1208.
Calamaro, Shira and Gaja Jarosz. 2015. Learning general phonological rules from distributional information: A computational model. *Cognitive Science* 39: 647–666
Calderone, Basilio, Nabil Hathouta, and Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. arXiv:2108.03968.
Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.
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.
Cutler, Anne, Andrea Weber, Roel Smits, and Nicole Cooper. 2004. Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America* 116:3668–78.
Czaykowska-Higgins, Ewa. 1988. Investigations into Polish morphology and phonology. PhD dissertation. MIT, Cambridge, MA.
Dai, Huteng. 2021. Learning nonlocal phonotactics in a strictly piecewise probabilistic phonotactic model. In *Proceedings of the Annual Meetings on Phonology*. Vol. 8.
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.
DelBusso, Natalie. 2020. Learning with properties: restrictiveness and typological structure. In *Proceedings of the Annual Meetings on Phonology*. Vol. 7.
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.
Do, Chuong B., and Serafim Batzoglou. 2008. What is the expectation maximization algorithm? *Nature Biotechnology* 26:897-899.

Dowd, Andrew. 2005. Surface base selection in Pengo. *Proceedings of the 24th West Coast Conference on Formal Linguistics*, ed. John Alderete et al., 105-111. Somerville, MA: Cascadilla Proceedings Project.

Dresher, B. E. 1981. On the learnability of abstract phonology. In C. L. Baker and John J. McCarthy (eds.), *The logical problem of language acquisition*, pp. 188–210. Cambridge: MA: MIT Press.

Dreyer, Markus and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 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, Armando Solar-Lezama, and Joshua B. Tenenbaum. 2015. Unsupervised learning by program synthesis. In *Advances in neural information processing systems*, pages 973–981.

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.

Ellison, T. Mark. 1994. Phonological derivation in Optimality Theory. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, 1007–1013. Kyoto.

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.

Fleischhacker, Heidi. 2005. Similarity in phonology: Evidence from reduplication and loan adaptation. Ph.D. dissertation, UCLA, Los Angeles, CA.

Foley, James. 1965. Spanish morphology. Ph.D. dissertation, MIT, Cambridge, MA.

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Stockholm workshop on variation within Optimality Theory*, 111–120. Stockholm: Stockholm University.

Gorecka, Alicja. 1988. Epenthesis and the coda constraints in Polish. Cambridge, MA: MIT, MS.

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 Michael Becker. 2013. Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language and Linguistic Theory* 31:735–765.

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*(2), pp.309-350.

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

Hayes, Bruce. 1995. On what to teach the undergraduates: Some changing orthodoxies in phonological theory. In Ik-Hwan Lee, ed., *Linguistics in the Morning Calm 3*, Hanshin, Seoul, pp. 59-77

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

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, Adam Jardine, and Huteng Dai. 2021. Learning underlying representations and input-strictly-local functions. *Proceedings of WCCFL 37*:143-151.

Hua, Wenyue, and Adam Jardine. 2021. Learning input strictly local functions from their composition. In Chandlee et al., eds., *Proceedings of the Fifteenth International Conference on Grammatical Inference*, PMLR 153:47-65.

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 Caroline Féry & Ruben van de Vijver (eds.) *The syllable in Optimality Theory*. Cambridge University Press. 271–303.

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

Jardine, Adam, and Kevin McMullin. 2017. Efficient learning of tier-based strictly k-local languages. In *Language and Automata Theory and Applications: 11th International Conference*, ed. by Frank Drewes, Carlos Martín-Vide, and Bianca Truthe, 64–76. Cham: Springer.

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

Jarosz, Gaja. 2008. Partial ranking and alternating vowels in Polish. In Rodney L. Edwards, Patrick J. Midtlyng, Colin L. Sprague, and Kjersti G. Stensrud, eds., *Proceedings of the Chicago Linguistic Society* 41, vol. 1. Chicago: CLS, 193-206.

Jarosz, Gaja. 2013. Learning with hidden structure in Optimality Theory and Harmonic Grammar: Beyond robust interpretive parsing. *Phonology* 30: 27-71.
Jarosz, Gaja. 2015. Expectation driven learning of phonology. Ms., University of Massachusetts.
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*, pages 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 the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 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. Ph.D. 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 Emmon Bach and Robert T. Harms (eds.) *Universals in linguistic theory*. New York: Holt, Rinehart & Winston. 170–202.
Kiparsky, Paul. 1971. Historical linguistics. In William Orr Dingwall (ed.) *A survey of linguistic science*. College Park: University of Maryland Linguistics Program. 576-642.
Kiparsky, Paul. 1973. Abstractness, opacity, and global rules. In Osamu Fujimura (ed.) *Three dimensions in linguistic theory*. Tokyo: TEC. 57–86.
Kiparsky, Paul. 1982. Lexical phonology and morphology. In I.-S. Yang (ed.), *Linguistics in the Morning Calm.* Seoul: Hanshin. 3-91.
Kirchner, Robert. 1996. Synchronic chain shifts in Optimality Theory. *Linguistic Inquiry* 27: 341-350.
Kisseberth, Charles. 1970. Vowel elision in Tonkawa and derivational constraints. In *Studies presented to Robert B. Lees by his students*, ed. Jerrold M. Sadock and Anthony L. Vanek, 109–137. Edmonton; Linguistic Research.

Khalifa, Salam, Sarah Payne, Jordan Kodner, Ellen Broselow, and Owen Rambow. 2023. A cautious generalization goes a long way: Learning morphophonological rules. *ACL (Volume 1: Long Papers)*, 1793-1805.

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 (to appear) Phonological reanalysis is guided is guided by markedness: the case of Malagasy weak stems. To appear in *Phonology*.

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.

Legendre, Geraldine, Yoshiro Miyata and Paul Smolensky. 1990. Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An Application. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 884–891. Mahwah, NJ: Lawrence Erlbaum Associates.

Liang, Kevin, Victoria Mateu, and Bruce Hayes (in preparation) A wug-test study of the Catalan consonant alternations. Ms., Dept. of Linguistics, UCLA.

Lightner, Theodore. 1965. Segmental phonology of modern standard Russian. Ph.D. 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*(2), pp.271-319.

McCarthy, John J. 2018. How to delete. *Perspectives on Arabic Linguistics XXX*. Available at: http://works.bepress.com/john_j_mccarthy/114/

McCarthy, John J., and Alan Prince. 1993. Generalized alignment. In *Yearbook of morphology 1993*, pp. 79-153. Dordrecht: Springer Netherlands.

McCarthy, John J. and Prince, Alan. 1995. Faithfulness and reduplicative identity. *Papers in Optimality Theory* 10. Department of Linguistics, University of Massachusetts.

McLachlan, Geoffery J., and Thriyambakam Krishnan. 1997. *The EM algorithm and extensions*. Wiley, New York.

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(1):61–66.

Mascaró, Joan. 1976. Catalan phonology and the phonological cycle. Ph.D. dissertation, MIT.

Meng, Xiao-Li, and Donald B. Rubin. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80:267–78.

Merchant, Nazarré. 2008. Discovering underlying forms: contrast pairs and ranking. PhD dissertation, Rutgers University.

Merchant, Nazarré, and Bruce Tesar. 2008. Learning underlying forms by searching restricted lexical subspaces. *Chicago Linguistic Society* 41:33–47.
Moore-Cantwell, Claire, and Robert Staubs. 2014. Modeling morphological subgeneralizations. in *Proceedings of the Annual Meeting on Phonology* 1.
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. *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: Iss. 1, Article 21.
Nyman, Alexandra, and Bruce Tesar. 2019. Determining underlying presence in the learning of grammars that allow insertion and deletion. *Glossa: A journal of general linguistics* 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, and Brandon Prickett. 2022. Typological gaps in iambic nonfinality correlate with learning difficulty. In *Proceedings of the Annual Meetings on Phonology*, vol. 9.
Pater, Joe, Robert Staubs, Karen Jesney, and Brian Cantwell Smith. 2012. Learning probabilities over underlying representations. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pp. 62-71.
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.
Pierrehumbert, Janet. 2006. The statistical basis of an unnatural alternation, in Louis Goldstein, Douglas Whalen and Catherine T. Best (eds.), *Laboratory Phonology* VIII. Berlin: Mouton de Gruyter 81-107.
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 Hucklebridge, S. and Nelson, M., editors, *Proceedings of NELS 48*, 283-290.
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. Ph.D. dissertation, Stony Brook University.

Richter, Caitlin. 2021. Alternation-sensitive phoneme learning: Implications for children's development and language change. Ph.D. dissertation, University of Pennsylvania.
Riggle, Jason. 2004. Generation, recognition, and learning in finite state Optimality Theory. Ph.D. 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.
Shilen, Alex and Colin Wilson. 2022. Learning input strictly local functions: comparing approaches with Catalan adjectives. In *Proceedings of the Society for Computation in Linguistics 2022*, 244–246. Association for Computational Linguistics.
Smolensky, Paul. 1996. The initial state and "richness of the base" in Optimality Theory. Technical report JHU-CogSci-96-4, Department of Cognitive Science, The Johns Hopkins University, Baltimore, Md. Rutgers Optimality Archive ROA-154.
Sommerstein, Alan. 1977. *Modern phonology*. London: Edward Arnold.
Steriade, Donca. 2000. Paradigm uniformity and the phonetics-phonology boundary. *Papers in Laboratory Phonology* 5, ed. by Michael Broe and Janet Pierrehumbert, 313-334. Cambridge: Cambridge University Press.
Steriade, Donca. 2001. Directional asymmetries in place assimilation: A perceptual account. *The role of speech perception in phonology*, ed. by Elizabeth Hume and Keith Johnson, 219-250. San Diego: Academic 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. 2004. Using inconsistency detection to overcome structural ambiguity. *Linguistic Inquiry* 35:219–253.
Tesar, Bruce. 2014. *Output-driven phonology: Theory and learning*. Cambridge: Cambridge University Press.
Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.
Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge: MIT Press.
Tesar, Bruce, John Alderete, Graham Horwood, Nazarré Merchant, Koichi Nishitani, and Alan Prince. 2003. Surgery in language learning. In *Proceedings of the Twenty-Second West Coast Conference on Formal Linguistics*, pp. 477-490.
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. 2018. Modeling morphological affixation with interpretable recurrent networks: sequential rebinding controlled by hierarchical attention. *CogSci 2018*: 2693–2698.

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.

Wilson, Colin, and Jane S. Y. Li. 2021. Were we there already? Applying Minimal Generalization to the SIGMORPHON-UniMorph shared task on cognitively plausible morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 283-291.

Wiemerslage, Adam, Arya McCarthy, Alexander Erdmann, Garrett Nicolai, Manex Agirrezabal, Miikka Silfverberg, Mans Hulden, and Katharina Kann. 2021. The SIGMORPHON 2021 Shared Task on Unsupervised Morphological Paradigm Clustering. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics.

Xu, Hongzhi, Jordan Kodner, Mitchell Marcus, and Charles Yang. 2020. Modeling morphological typology for unsupervised learning of language morphology. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6672–6681.

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. Ph.D. dissertation, UCLA.

Zuraw, Kie. 2000. Patterned exceptions in phonology. Ph.D. 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.