A learnability argument for constraints on underlying representations*

Ezer Rasin and Roni Katzir

October 23, 2014

1 Where are phonological generalizations captured?

As noted by Halle (1962) and Chomsky and Halle (1965), speakers of a language judge some nonce forms as nonexistent but possible – that is, as *accidental gaps* – and other nonce forms as nonexistent and impossible – that is, as *systematic gaps*. In English, for example, forms such as gæ̃n and gæd are accidental gaps, while gæ̃d and gæn are systematic gaps. Capturing this distinction in speaker's judgments is a central task of phonological theory. Since the judgments of speakers regarding nonce forms differ between languages, this task also involves accounting for how the relevant knowledge is acquired.

Early generative approaches accounted for systematic gaps through a combination of two factors: constraints on underlying representations in the lexicon;[1] and phonological rules. In the example above, an early generative account might use the following constraint on URs as the basis for capturing the distribution of nasalization in English:

(1) Constraint on URs in English: No nasal vowels in the lexicon

With (1) in place, a phonological rule can complete the picture by nasalizing prenasal vowels. The accidental gæ̃d and gæd could be added to English with the URs /gæn/ and /gæd/; the nasalizing rule would then turn the former into its surface form. For gæ̃d and gæn the situation is different: since nasalization is not stored in the lexicon of English, the nasalization in gæ̃d must follow from rule application, but the nasalization rule does not apply to vowels that are not prenasal; for gæn, on the other hand, obligatory nasalization would ensure that this surface form cannot appear. Both gaps are thus correctly treated as systematic.

Contrasting with this view, Optimality Theory (Prince and Smolensky, 1993, OT) has been guided by the idea that phonological generalizations are captured not in the lexicon but rather on the surface or in the mapping from URs to surface forms. In the

*Acknowldgments: To be added.

[1] Halle (1959, 1962) proposed to capture the relevant generalizations through rules that apply to URs. Stanley (1967) argued that these should be constraints rather than rules. In the generative tradition these became known as morpheme-structure constraints. We use constraints on URs as a cover term for rules or constraints of this kind.

example above, markedness constraints – as toy examples, *ãd and *an – would penalize nasal non-prenasal vowels and oral prenasal ones. Ranking these constraints higher than the relevant faithfulness constraints would ensure that even URs with inappropriately nasalized vowels will surface correctly, thus correctly ruling out gæn and gæ̃d as systematic gaps. The accidental gaps gæ̃n and gæd, on the other hand, can be added to the lexicon with URs that are identical to the surface forms.

Differently from rule-based phonology, then, where capturing the distinction between accidental and systematic gaps makes central use of constraints on URs such as (1) above, OT can capture the distinction without such constraints.[2] The ability of OT to distinguish between accidental and systematic gaps without recourse to constraints on URs suggests that perhaps constraints on URs are *never* used. This stronger view is often referred to as Richness of the Base, and it is a central tenet of OT:

(2) Richness of the Base (ROTB; Prince and Smolensky 1993, p. 191,Smolensky 1996, p. 3):
   a. All systematic language variation is in the ranking of the constraints.
   b. In particular, there are no language-specific constraints on URs.

Our goal in this note is to re-open the question of whether OT requires constraints on URs and offer a learnability argument supporting an affirmative answer, thus arguing against ROTB. We start, in section 2, by examining the extant literature on learning in OT. The relevant works adopt the representational principle of ROTB, often in combination with two learning principles that are inspired by it. One of these learning principles is sometimes also referred to as ROTB, but to avoid confusion we will call it by a different name, IMAGINE, and reserve the term ROTB for the representational principle in (2). The second learning principle is known as Lexicon Optimization (LO). We argue that approaches based on IMAGINE and LO are untenable: they all overgeneralize (by treating some systematic gaps as accidental), undergeneralize (by treating some accidental gaps as systematic), or both. In section 3 we discuss a different approach to learning, compression-based learning, that does not assume IMAGINE or LO and that is the only approach currently available that can handle the data in principle without over- or undergeneralization. In section 4 we show that compression-based learning learns certain naturally-occurring patterns but crucially only if it rejects ROTB and employs language-specific constraints on URs.

2 IMAGINE and LO, and why they should be abandoned

As mentioned, the representational principle of ROTB is often combined with two learning principles, IMAGINE and LO. Our main focus in this paper is ROTB, and we

[2] In certain cases, constraints on URs in rule-based phonology seem to conspire with phonological rules to enforce what looks like a unitary surface effect (see Chomsky and Halle 1968, p. 382). This fragmentation of explanation – the *duplication problem* of Kenstowicz and Kisseberth (1977, p. 136) – has been taken to be a challenge to rule-based phonology and an argument in favor of OT. Below we will argue that constraints on URs should be brought back into phonological theory. In the cases we will discuss, however, we will see no fragmentation between such constraints and other mechanisms, and we will not attempt to re-evaluate the duplication problem.

will present a learnability argument against it. Before we do that, though, we will need to isolate it from the two learning principles, which is what we do in the present section. Specifically, we will argue that both learning principles are untenable.

2.1 ROTB, IMAGINE, and LO

The first learning principle, IMAGINE (Smolensky 1996, Tesar and Smolensky 1998), encourages the learner to look for a constraint ranking that yields the observed data not just given one lexicon but given many other hypothetical URs:[3]

(3) IMAGINE: Prefer constraint rankings that make the observed data typical not just given a particular lexicon but with respect to every imaginable UR.

The motivation for IMAGINE, as discussed by Smolensky (1996) and Tesar and Smolensky (1998) (attributing the original idea to Alan Prince), is the need to make the grammar restrictive. Considering the observed data in view of just one lexicon is too easy: the learner can get away with positing URs that are identical to the surface forms along with the identity mapping, implemented by ranking faithfulness above markedness ($F \gg M$). This would allow the learner to ignore the entire sound pattern of the language. In English, for example, this approach would treat the absence of clicks or of rt onsets as accidents of the lexicon. Nothing will force the learner to state these generalizations in the ranking of the constraints. Speakers of English know that clicks and rt onsets are impossible in their language – mip and trɛiv, for example, are possible English words, but ʘip and rtɛiv are not – so the lazy $F \gg M$ option must be blocked.

IMAGINE has been offered as the remedy. By considering not just one lexicon but all forms, the learner now has to ensure that clicks and rt onsets will never surface, not even if they are present underlyingly. This, in turn, forces the learner to state the relevant generalizations within the ranking of the constraints. In English, for example, constraints against clicks and rt onsets will now be ranked above faithfulness constraints.

The letter of IMAGINE has proven to be difficult to implement directly within standard OT, leading to a variety of approaches that aim at capturing the spirit of the principle: an initial ranking of M≫F (Smolensky 1996, Tesar and Smolensky 1998) from which a search for a consistent ranking begins; a more sustained bias for M≫F (Hayes 2004, Prince and Tesar 2004) throughout the search for a consistent ranking; Maximum Likelihood (Jarosz 2006); and others. While differing among themselves in various ways, these implementations pattern together with respect to the arguments that we will discuss below. We will thus group them together and refer to them as IMAGINE-based learners.

The second learning principle, Lexicon Optimization (LO; Prince and Smolensky 1993, p. 209, Inkelas 1995, Smolensky 1996), encourages the learner to do as little work as possible in guessing the URs of surface forms. Given a surface form and a

[3] As mentioned, the literature sometimes uses ROTB to refer to IMAGINE. The two are very different creatures, however, and though we will end up concluding that both are incorrect, we will be careful to keep them separate in our discussion.

particular constraint ranking, there can be many different URs that would yield that surface form. Of these potential URs, LO instructs the learner to guess the one with the least significant violations of constraints – the UR that yields the most *harmonic* mapping, to use common OT parlance. When a morpheme has different surface forms as part of a paradigm with alternations, Inkelas (1995) proposes an extension of LO that considers the entire paradigm and guesses a UR that is most harmonic with respect to the entire paradigm (specifically, by incurring the fewest violations of most highly ranked constraints).

(4) Lexicon Optimization: In the face of multiple potential URs that yield the same surface form, choose the most harmonic one.[4]

LO tells the learner to choose URs that are as close as possible to the observed surface forms. For a morpheme that does not participate in an alternation, a UR that is identical to the surface form will always be maximally harmonic: URs are not subject to any markedness constraints in OT, and identity between UR and surface form can incur no faithfulness violation. An LO learner will therefore be justified in choosing the surface form as its own UR whenever no alternation is involved.[5]

IMAGINE and LO are natural companions to ROTB. ROTB allows any possible UR to be stored in principle, even URs that are very different from anything seen so far in the language. IMAGINE takes advantage of this and reasons hypothetically about different URs to obtain a restrictive constraint ranking. And LO takes advantage of the same property to store URs with as little modification as is necessary. Despite this naturalness, IMAGINE and LO fail as learning principles, as we now show.

2.2 IMAGINE overfits the data: the problem of accidental gaps

IMAGINE-based learners seek to capture as many generalizations as possible in the constraint ranking. This leads to an overly restrictive grammar, where accidental gaps are treated as systematic and restrictiveness is limited only by the possible representations provided by UG. We demonstrate this with two examples.

2.2.1 A simple example

Suppose UG makes available all markedness constraints of the general form $^*X_1 \ldots X_j$ (maybe up to a certain length). In the data, many sequences of this form will usually be missing (in particular, any such sequence that is longer than the longest word in the data). IMAGINE-based approaches will lead to the ranking of any constraint of the form $^*X_1 \ldots X_j$ corresponding to a gap above all faithfulness constraints. Halle (1962)'s example of θo͡ʊd, a nonexistent but possible English word, makes this point using a

[4] Prince and Smolensky (1993, pp. 219–210) also consider a different condition, Minimal Lexical Redundancy, and also a constraint called *SPEC that penalizes all underlying material.

[5] Depending on the choice of faithfulness constraints and of representations, it is sometimes possible for a non-identical UR to be as harmonic as an identical one. For example, if no faithfulness constraint penalizes adding a feature to an underspecified form, /na/, with a nasal that is unspecified for place of articulation, and /ma/, where place is specified as labial, are both optimal URs for a non-alternating [ma]. See Krämer (2005) for discussion. This observation does not affect the arguments against LO presented below.

short sequence of segments: if an IMAGINE-learner has a constraint such as *θo͡ʊd, it will rank the markedness constraint (which is never violated in the data) high enough to make it treat the gap, incorrectly, as systematic rather than accidental.

A possible response to Halle's example is that it simply teaches us that *θo͡ʊd ∉ CON. We are not aware of a principled reason to exclude *θo͡ʊd from CON, but to clarify this matter – and to tighten the argument from θo͡ʊd against IMAGINE – we would need to find another language in which θo͡ʊd is an impossible word (and not because of other, independent constraints). We are not currently familiar with a language that can make this point using θo͡ʊd, but we believe that the same point can be made using a slightly more complex case, to which we now turn.

2.2.2 A more complex example

Constraints of the form $^{*}sC_iVC_i$ are active in English, e.g., $^{*}skVk$, $^{*}slVl$, $^{*}sNVN$ where N is a nasal consonant. In Hebrew, $sNVN$ and $skVk$ are accidental gaps: (a) sN and sK are permissible onset clusters (*snif* 'branch', *skira* 'survey'); (b) NVN and kVk are attested word endings (*minun* 'dosage', *zakuk* 'needs'); (c) there are also a few words of the form $sCVC$ (*sxax* 'cover', *slil* 'coil'); (d) but no word of the form $sNVN$, $skVk$. An IMAGINE-based learner will rank the relevant markedness constraints – including those that correspond to accidental gaps – above all faithfulness constraints, leading to the gaps being treated incorrectly as systematic.[6,7]

2.2.3 A note on phonotactic grammars

Before leaving IMAGINE-based learners, we wish to briefly consider a different constraint-based approach to the challenge of accidental and systematic gaps. This alternative approach is the phonotactic grammars and learning procedure of Hayes and Wilson (2008). While their approach is explicitly non-OT, it could be used to complement an OT grammar – in fact, Hayes and Wilson themselves suggest such an architecture (p. 424). In a combined architecture, Hayes and Wilson's phonotactic component might take care of separating accidental from non-accidental gaps, after which IMAGINE can take over and acquire the remaining phonological knowledge.

Unfortunately, this combined architecture will probably not work, since Hayes and Wilson (2008)'s approach suffers from the exact same problem of overfitting as IMAGINE-based learners. Their algorithm (p. 394) combines heuristics for selecting con-

[6] A more realistic (but possibly less relevant) example is due to Zimmer (1969), who notes that Turkish speakers apply only some of these generalizations (e.g., vowel harmony) and not others (e.g., labial attraction) to novel words. See Becker et al. (2011) for additional evidence supporting this point. Though: this line of work focuses on statistical tendencies rather than absolute generalizations.

[7] Adam Albright (p.c.) raises the following direction for an IMAGINE-based account of sC_iVC_i, following Coetzee (2008). For Coetzee, the $^{*}sC_iVC_i$ constraints are not primitive but rather obtained through constraint conjunction of a general $^{*}sC$ and specific non-repetition constraints such as $^{*}tVt$, $^{*}kVk$, and $^{*}pVp$. The choice to form the relevant constraint conjunctions is moreover taken to follow from the statistics of the violation of the constituent constraints in the data. It is conceivable, then, that English – but not Hebrew – will have statistics that induce the formation of the relevant constraint conjunctions, which could account for the difference between the two languages with respect to sC_1VC_1-shaped gaps. Since this direction assumes various mechanisms that are as yet unspecified (in particular, the criterion for forming constraint conjunction), we do not discuss it further here.

straints for a MaxEnt grammar with weight training aimed at maximizing the likelihood of the data. Maximizing the likelihood will of course try to overfit the data, so the only hope of the learner is that the heuristics for constraint selection will prevent it from adding constraints like *θo͡ʊd. The heuristics, however, all but guarantee that constraints like *θo͡ʊd will be added. The algorithm proceeds from high precision (that is, few exceptions) to lower precision, adding the constraints at that level of precision (by decreasing order of generality, according to their two-step definition of generality, pp. 393–4). *θo͡ʊd has no exceptions in English, so if it can be represented and if the search is not stopped too soon, it will be added.[8] Not surprisingly, perhaps, Hayes and Wilson's algorithm acquires both meaningful-looking patterns and what seem to be arbitrary patterns, as the authors themselves note (pp. 419–20). Hayes and White (2013) run the same learner on English data and provide experimental evidence that only the meaningful-looking patterns are active in speakers' knowledge; it remains to be seen whether this distinction corresponds to that between overfitted and non-overfitted patterns based on a more balanced learning criterion (Bayes, MDL, etc.). As far as we can tell, then, using Hayes and Wilson (2008)'s approach for phonotactics alongside – or instead of – an IMAGINE-based OT learner provides no defense against overfitting.

2.3 LO

2.3.1 Learners should not optimize

Recall that, in the absence of alternations, LO will always be satisfied by a UR that is identical to the surface form. Alderete and Tesar (2002) note that this makes LO an obstacle to learning phonological patterns such as the stress patterns in Mohawk, Selayarese, and Yimas. In these languages, the URs differ from their corresponding surface forms with respect to stress assignment. Significantly, however, the learner cannot rely on alternations to point to the discrepancies in these languages. This amounts to a direct challenge to LO.

To illustrate, consider the schematic stress typology in the following table, and in particular the stress pattern in Language *B*:

Name	Type	Lexicon							
Lg. A	Lexical stress	pákat	pakát	pákit	pakít	píkat	pikát	píkit	pikít
Lg. B	Stress-epenthesis		pakát	pákit	pakít		pikát	píkit	pikít
Lg. C	Final stress		pakát		pakít		pikát		pikít

Languages *A* and *C* are straightforward: the former has lexical stress and the latter final stress, and in both cases the URs could be identical to the surface forms (for Language *A*, this will have to be the case; for Language *C* this is assumed to be the case by adherents of LO, but considerations such as transparency may lead to URs

[8] Concretely, Halle's *θo͡ʊd happens to be just a little bit too long for Hayes and Wilson's representations: they admit segmental constraints of up to three feature vectors, but one of those vectors has to be degenerate; we take it, however, that this restriction is imposed to ensure reasonably fast convergence on the original hardware used by the authors and not due to more principled considerations. Moreover, we can replicate the argument with θom, a slight modification of Halle's example that is also a possible but not actual English word. Differently from θo͡ʊd, for which θʌd ('*thud*') is an actual English word, no vocalic variant of θo͡ʊb is an English word. Consequently, Hayes and Wilson (2008)'s learner has no principled barrier to learning the problematic constraint *θVb.

that are underspecified for stress assignment, as discussed above). Language *B* is more problematic: it has final stress in general, but when the final vowel is [i], stress can be penultimate. A standard analysis would take *B* to have final stress, treating unstressed final [i] as epenthetic. Arguing for a stress-epenthesis account of *B* would require, among other things, testing the prediction that speakers reject as impossible words forms such as *pákat* with penultimate stress and without a final [i]. Assuming that this prediction is borne out, *B* constitutes a major challenge for IMAGINE-LO learners. The relevant URs do not have the final [i] appearing on the surface, and there is no alternation to provide the crucial evidence for non-identity. Learners based on IMAGINE and LO thus converge on an overly inclusive analysis of *B* as a lexical-stress language (e.g., predicting that an impossible **pákat* would be a possible word).[9]

Again, a phonotactic front-end along the lines of Hayes and Wilson (2008) will be of no help. In the present case, such a component might miss the relevant generalization about final stress since it is not surface-true (whether it will or will not depends on the statistical distribution of epenthesis in the language and on the specific parameters of the phonotactic learner). And even if it does acquire the relevant constraint, it will be demoted during the IMAGINE-LO stage, as discussed above, and language *B* will be incorrectly taken to have lexical stress

2.3.2 Learners do not optimize

There is no clear evidence to date regarding the inferences that children make regarding URs during phonological acquisition. To the extent that these URs can be probed at later stages, however, they are the exact opposite of what LO predicts. This point has been made by Vaux (2005) and Nevins and Vaux (2007), who present a variety of cases in which URs differ from surface forms even in the absence of supporting alternations.[10] For example, Nevins and Vaux note that when speakers of languages like German and Russian that have final devoicing are presented with a nonce word such as mik, they will sometimes posit the unfaithful UR mig. This is entirely surprising from the perspective of LO. It is equally puzzling from the perspective of other absolute principles of UR induction, such as McCarthy (2005)'s *Free Ride Principle*, which tells learners who have observed an alternation and used it to obtain a non-faithful UR in some cases to do the same in all cases: for such learners, mig will be posited as the UR for mik not just sometimes but always. Again, the empirical pattern is more nuanced, showing unfaithful URs posited in some cases but not in all of them.[11]

In the mik~mig case, one can try to devise an account that relies on the devoicing of other words in order to inform the unfaithful UR posited by speakers. As Nevins and Vaux note, however, speakers sometimes posit unfaithful URs not just when some

[9] It might be possible to construct an account of Language *B* that relies on a phonetic difference between regular and epenthetic vowels (cf. Gouskova and Hall 2009 and Hall 2013). For Yimas, on which Language *B* is modeled, we know of no basis for an alternative explanation along phonetic lines, but we note that settling this matter would require a more thorough investigation than we are offering here.

[10] See also Harrison and Kaun (2000, 2001), Ernestus and Baayen (2003), and Krämer (2012).

[11] Nevins and Vaux suggest that the proportion of unfaithful URs that speakers posit will correspond not to an absolute rule such as LO but rather to the frequency of the relevant alternation (in this case, final devoicing) in the language.

related form participates in an alternation but also when such information is unavailable. They consider the case of rhotics in Spanish, which can be realized as r or ɾ word-medially but only as r word-initially. When induced to move a word-initial r to a word-medial position as part of a language game, speakers sometimes realized it as r, in line with a faithful UR, but sometimes as ɾ, suggesting an unfaithful UR.

3 Compression-based learning

As we have just seen, IMAGINE and LO face significant empirical challenges while arguments supporting them are thin. An impediment to abandoning IMAGINE and LO, even in the face of these challenges, has been the absence of a plausible algorithm that would work in their absence. To our knowledge, there is exactly one approach in the literature that offers a handle on both the over-generalization problem and the under-generalization problem. This approach, which we will refer to as *compression-based learning*, aims at balancing the complexity (or probability) of the grammar with that of the grammar's account of the data. Specific implementations of compression-based learning sometimes follow the principle of Minimum-Description Length (MDL; Rissanen 1978) and sometimes the closely-related idea of Bayesian reasoning. The roots of both are in the pioneering work of Solomonoff (1964), and other early work includes Wallace and Boulton (1968) and Horning (1969). Compression-based learning has been used for grammar induction in the works of Berwick (1982), Rissanen and Ristad (1994), Stolcke (1994), Brent and Cartwright (1996), Chen (1996), Grünwald (1996), de Marcken (1996), Osborne and Briscoe (1997), Brent (1999), Clark (2001), Goldsmith (2001), Onnis et al. (2002), Zuidema (2003), Dowman (2007), and Chang (2008), among others.

Recently, we have proposed a compression-based learner for OT in Rasin and Katzir (2014), and it is this learner that we will use in our argument against ROTB and in favor of constraints on URs. Following the principle of MDL, the learner attempts to minimize the overall description of the data, measured in bits. The overall description is broken down into G, the encoding of the grammar (which, for OT, includes both the lexicon and the constraints), and $D|G$, the description of the data D given the grammar. The combination of grammar and data is schematized in Figure 1 (modified from Rasin and Katzir 2014).

$$\underbrace{\underbrace{0101011010101001010}_{Lexicon}\,\underbrace{10101010010}_{Constraints}}_{G}\,\underbrace{10100010110101}_{D|G}$$

Figure 1: Schematic view of an OT grammar and the data it encodes. The grammar G consists of both lexicon and constraints. The data D are represented not directly but as encoded by G. The overall description of the data is the combination of G and $D|G$.

As discussed in Rasin and Katzir (2014), the length of G, $|G|$, corresponds to the informal notion of economy, familiar from the evaluation metric of Chomsky and Halle

(1968): a grammar that requires fewer bits to encode is generally a simpler, less stipulative grammar. Meanwhile, the length of $D|G$, $|D|G|$, corresponds to restrictiveness, the goal of IMAGINE: a grammar that requires fewer bits to encode the data is a grammar that considers the data typical and deviations from the data surprising. Crucially, compression-based learning dictates the simultaneous balancing of economy and restrictiveness: the learner attempts to minimize the overall description length, $|G| + |D|G|$, rather than prioritizing one factor over the other. This balancing of economy and restrictiveness is exactly what protects compression-based learners from both over-generalization and under-generalization.

Compression-based learning is a very general approach.[12] It is not tied to the specifics of OT, and it is independent of the question of ROTB. This generality will allow us to combine compression-based learning with two versions of OT – one that assumes ROTB and one that does not – and compare their predictions on two patterns that occur in natural languages. We show that the version without ROTB successfully learns the patterns, but the version that assumes it fails. The failure of the version while assuming ROTB will not be accidental: as we will see, constraints on URs are indispensable for compression-based learning of these patterns. Since compression-based learning is the only approach currently available that addresses the challenges of over-generalization and under-generalization, this will amount to an argument against ROTB.

4 The argument against ROTB

4.1 Simulation I: Aspiration in English

Our first demonstration that compression-based learning must abandon ROTB comes from a simplified version of aspiration in English. In this dataset, stops are aspirated exactly when they are prevocalic.[13] A compression-based learner can learn correctly that prevocalic stops are aspirated by removing instances of aspiration from the lexicon (e.g., /*kat*/ → [$k^h at$]) and ranking the constraints accordingly. Crucially, however, it will only learn to block $^*at^h$ and $^*k^h ik^h t$ if it can systematically ban aspiration from the lexicon, thus implementing a constraint on URs along the lines of (1) above:

(5) CONSTRAINT ON URS IN ENGLISH: No aspirated consonants in the lexicon

To see why a constraint-based learner needs something like (5) to block aspiration outside of prevocal environments, suppose the learner allowed aspiration to be represented underlyingly in principle. This would mean that the final grammar would have to rule out forms such as at^h, $kik^h t$, etc. through the input-output mapping. But the learner has no reason to learn to block such forms through the input-output mapping: in the actual lexicon, as just mentioned, the URs are stored without aspiration;

[12] See Katzir (2014) for an argument that the MDL version of compression-based learning is available to the learner by virtue of having a theory of competence, which makes this approach a fully general null hypothesis.

[13] The aspiration pattern could in principle be handled by IMAGINE-LO learners or by phonotactic learners. This, however, would be irrelevant to our argument: as discussed above, such learners fail by both over- and under-generalizing and so are out of the game by now.

and without instances of underlying aspiration, a constraint that ensures that aspiration does not surface in illicit positions will serve no compressional purpose. In particular, such a constraint will not make the data more likely (or easier to describe) given the grammar; consequently, it will fail to make the hypothesis preferred over an alternative hypothesis that does not ban instances of aspiration from surfacing in illicit positions.

On the other hand, a constraint on URs such as (5) has the potential to add compressional value. In particular, suppose that (5) is implemented by removing aspiration from the inventory of primitives used for URs. All things being equal, removing a possible segment from the underlying inventory makes it slightly easier to specify the remaining segments, some of which may now cost fewer bits than before. Consequently, the lexicon will now be encoded with fewer bits, thus providing compressional justification for adopting (5). And as in rule-based phonology, adopting (5) ensures that surface forms like at^h and kik^ht will be blocked: due to (5), /at^h/ and /kik^ht/ are no longer possible URs; and such URs are the only potential source for surface aspiration in inappropriate contexts. In other words, the impossibility of storing aspiration in the lexicon, with its compressional justification discussed above, means that the learner has correctly learned to block bad aspiration.

The simulation results summarized here support this conclusion.[14] The data available to the learner, with a sample in (6), are generated from a small segmental inventory subject to the condition that aspiration can only appear prevocalically; aspiration is expressed as a separate segment, [^{h}]. In its initial state, summarized in (6b), the learner allows all segments, including aspiration, to appear in the lexicon. The markedness constraint $^*[+stop][-cons]$, which penalizes unaspirated prevocalic stops is initially outranked by the two faithfulness constraints, thus failing to enforce aspiration in the relevant position. At the end of the simulation, summarized in (6c), $^*[+stop][-cons]$ outranks the faithfulness constraints, thus enforcing prevocalic aspiration. More importantly, the final segmental inventory is without [^{h}]: aspiration has been eliminated from the lexicon – a constraint on URs – ensuring that inappropriately aspirated segments will not be possible words in the language. When the learner is not allowed to eliminate aspiration from the lexicon, this last step cannot take place.

(6) a. **Data:** $\{k^hik, t^hatk, k^hak^hiat, \dots\}$

b. **Initial state:**
$\Sigma = \{a, i, t, k, {}^h\}$; Lex: $\{k^hik, t^hatk, k^hak^hiat, \dots\}$
Con: Faith≫Max[+*asp*]≫$^*[+stop][-cons]$

c. **Final state:**
$\Sigma = \{a, i, t, k\}$ (no [^{h}]); Lex: $\{kik, tatk, kakiat, \dots\}$
Con: $^*[+stop][-cons]$≫Faith≫Max[+*asp*]

4.2 Simulation II: Yimas stress-epenthesis interaction

Our second demonstration that compression-based learning should abandon ROTB comes from a simplified version of the interaction of stress and epenthesis in Yimas, a pattern we have already seen in 2.3.1 above. Recall that stress in Yimas is grammatical,

[14] See Rasin and Katzir (2014) for a detailed presentation of the learner and various simulations.

but predicting its exact location is complicated by vowel epenthesis. The compression-based learner succeeds in learning the stress pattern only if it can systematically ban underlying stress from the lexicon.

In the dataset summarized in (8a), stress in bisyllabic words is initial but can be second if the first vowel is [i]; as with the aspiration dataset, stress is represented as a separate segment, [′]. The interaction between stress and epenthesis is captured through the relative ranking of several constraints, central among them HeadDep, which penalizes stress on epenthetic vowels; MainLeft, which penalizes stress shifts from the first syllable; and $^{*}CC$, which penalizes consonant clusters (and can thus justify vowel epenthesis).[15] In the initial state, summarized in (8b), stress is still part of the segmental inventory in which the lexicon is written, and the constraint ranking fails to capture the basic pattern.[16] Compression leads to a short lexicon where stress and the relevant instances of [i] are systematically absent and the grammar inserts them in the right positions. Removing stress from the alphabet, as stated in (7), is driven by further compression – as with aspiration in English, the remaining underlying elements become slightly easier to specify, leading to compressional gains in the lexicon – and ensures that ungrammatical outputs (e.g., **katú*) are blocked. As with the aspiration dataset, preventing the learner from eliminating stress from the lexicon makes this last step impossible.

(7) Constraint on URs in Yimas: No stress marking in the lexicon

(8) a. **Data:** {*tí*, *púk*, *kátu*, *kúit*, *píkat*, *tipú*, *kipúk*}

b. **Initial state:**
$\Sigma = \{t, p, k, a, i, u, ′\}$
Lex: {*tí*, *púk*, *kátu*, *kúit*, *píkat*, *tipú*, *kipúk*}
Con: Faith≫Dep≫MainLeft≫$^{*}CC$≫HeadDep (simplified)

c. **Final state:**
$\Sigma = \{t, p, k, a, i, u\}$ (no [′])
Lex: {*ti*, *puk*, *katu*, *kuit*, *pikat*, *tpu*, *kpuk*}
Con: HeadDep≫$^{*}CC$≫Dep≫MainLeft≫Faith

5 Discussion

OT dispensed with constraints on URs for reasons of theoretical simplicity: a single-component architecture seemed more appealing than a dual-component one; output constraints unified constraints on URs and the input-output mapping. The present work brings learnability to bear on the question of whether constraints on URs are needed – specifically, to argue that they are. We first looked at Imagine and LO, the two learning principles that are used in the literature on learning in OT, and noted that they both

[15] The constraints used in the following simulation are taken from the literature (see, e.g., Alderete and Tesar 2002 and Jarosz 2009). We do not wish to defend this choice here, only to show how representations used in OT accounts of stress-epenthesis interaction can be learned using compression, but only if we admit constraints on URs.

[16] Specifically, we chose an initial ranking that is far away from the target one: faithfulness outranks markedness, and the relative ranking of the markedness constraints is the opposite of the correct ranking.

over-generalize and under-generalize. We pointed out that the only learning framework currently available that can handle learning in OT without running into these problems is compression-based learning. Since compression-based learning is a fully general approach, we could combine it with two OT architectures – one with CURs and one without – and use the predictions to choose between the two. We showed that learning empirically-attested patterns such as aspiration in English and stress-epenthesis interaction in Yimas requires CURs. Since compression-based learning is both cognitively plausible and the only current working approach for OT learning, this amounted to an argument in favor of constraints on URs and against the OT principle of ROTB.

References

Alderete, John, and Bruce Tesar. 2002. Learning covert phonological interaction: an analysis of the problem posed by the interaction of stress and epenthesis. Technical Report RuCCS-TR-72, Rutgers Center for Cognitive Science, Piscataway, NJ. ROA 543.

Becker, Michael, Nihan Ketrez, and Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87:84–125.

Berwick, Robert C. 1982. Locality principles and the acquisition of syntactic knowledge. Doctoral Dissertation, MIT, Cambridge, MA.

Brent, Michael. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Computational Linguistics* 34:71–105.

Brent, Michael, and T. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61:93–125.

Chang, Nancy Chih-Lin. 2008. Constructing grammar: A computational model of the emergence of early constructions. Doctoral Dissertation, EECS Department, University of California, Berkeley, Berkeley, CA.

Chen, Stanley. 1996. Building probabilistic models for natural language. Doctoral Dissertation, Harvard University, Cambridge, MA.

Chomsky, Noam, and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1:97–138.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row Publishers.

Clark, Alexander. 2001. Unsupervised language acquisition: Theory and practice. Doctoral Dissertation, University of Sussex.

Coetzee, Andries W. 2008. Grammaticality and ungrammaticality in phonology. *Language* 84:218–257.

Dowman, Mike. 2007. Minimum description length as a solution to the problem of generalization in syntactic theory. Ms., University of Tokyo, Under review.

Ernestus, Mirjam, and R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79:5–38.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:153–198.

Gouskova, Maria, and Nancy Hall. 2009. Acoustics of epenthetic vowels in Lebanese

Arabic. *Phonological Argumentation. Essays on evidence and motivation. London: Equinox* .

Grünwald, Peter. 1996. A minimum description length approach to grammar inference. In *Connectionist, statistical and symbolic approaches to learning for natural language processing*, ed. G. S. S. Wermter and E. Riloff, Springer Lecture Notes in Artificial Intelligence, 203–216. Springer.

Hall, Nancy. 2013. Acoustic differences between lexical and epenthetic vowels in Lebanese Arabic. *Journal of Phonetics* 41:133–143.

Halle, Morris. 1959. *The sound pattern of Russian*. Walter de Gruyter.

Halle, Morris. 1962. Phonology in generative grammar. *Word* 18:54–72.

Harrison, K. David, and Abigail Kaun. 2000. Pattern-responsive lexicon optimization. In *Proceedings of the North East Linguistic Society*, ed. Masako Hirotani, Andries Coetzee, Nancy Hall, and Ji-yung Kim, 327–340. Rutgers University: Graduate Linguistic Student Association.

Harrison, K. David, and Abigail Kaun. 2001. Patterns, pervasive patterns and feature specification. *Distinctive Feature Theory. Mouton de Gruyter, Berlin* 211–236.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: The early stages. In *Constraints in phonological acquisition*, ed. René Kager, Joe Pater, and Wim Zonneveld, 158–203. Cambridge, UK: Cambridge University Press.

Hayes, Bruce, and James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44:45–75.

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

Horning, James. 1969. A study of grammatical inference. Doctoral Dissertation, Stanford.

Inkelas, Sharon. 1995. The consequences of optimization for underspecification. In *Proceedings of the North East Linguistic Society 25*, ed. Jill Beckman, 287–302. University of Pennsylvania: Graduate Linguistic Student Association.

Jarosz, Gaja. 2006. Richness of the base and probabilistic unsupervised learning in Optimality Theory. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL 2006*, 50–59.

Jarosz, Gaja. 2009. Restrictiveness in phonological grammar and lexicon learning. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 43, 125–139. Chicago Linguistic Society.

Katzir, Roni. 2014. A cognitively plausible model for grammar induction. To appear in *Journal of Language Modelling*, Available at http://ling.auf.net/lingbuzz/001933, September 2014.

Kenstowicz, Michael, and Charles Kisseberth. 1977. *Topics in phonological theory*. Academic Press.

Krämer, Martin. 2005. Optimal underlying representations. In *Proceedings of NELS*, volume 35, 351–365.

Krämer, Martin. 2012. *Underlying representations*. Cambridge University Press.

de Marcken, Carl. 1996. Unsupervised language acquisition. Doctoral Dissertation, MIT, Cambridge, MA.

McCarthy, John J. 2005. Taking a free ride in morphophonemic learning. *Catalan Journal of Linguistics* 4:19–56.

Nevins, Andrew, and Bert Vaux. 2007. Underlying representations that do not minimize grammatical violations. In *Freedom of analysis?*, ed. Sylvia Blaho, Patrik Bye, and Martin Krämer, 35–61. Mouton de Gruyter.

Onnis, Luca, Matthew Roberts, and Nick Chater. 2002. Simplicity: A cure for overgeneralization in language acquisition? In *Proceedings of the 24th Annual Conference of the Cognitive Society*, ed. W. D. Gray and C. D. Shunn. London.

Osborne, Miles, and Ted Briscoe. 1997. Learning stochastic categorial grammars. In *Proceedings of CoNLL*, 80–87.

Prince, Alan, and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, Center for Cognitive Science.

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Constraints in phonological acquisition*, ed. René Kager, Joe Pater, and Wim Zonneveld, 245–291. Cambridge University Press.

Rasin, Ezer, and Roni Katzir. 2014. On evaluation metrics in Optimality Theory. Under review for *Linguistic Inquiry*, Available at http://ling.auf.net/lingbuzz/001934, February 2014.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Rissanen, Jorma, and Eric Sven Ristad. 1994. Language acquisition in the MDL framework. In *Language computations: DIMACS Workshop on Human Language, March 20-22, 1992*, 149. Amer Mathematical Society.

Smolensky, Paul. 1996. The initial state and 'richness of the base' in Optimality Theory. Technical Report JHU-CogSci-96-4, Johns Hopkins University.

Solomonoff, Ray J. 1964. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.

Stanley, Richard. 1967. Redundancy rules in phonology. *Language* 43:393–436.

Stolcke, Andreas. 1994. Bayesian learning of probabilistic language models. Doctoral Dissertation, University of California at Berkeley, Berkeley, California.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Vaux, Bert. 2005. Formal and empirical arguments for morpheme structure constraints. Talk given at LSA Annual Meeting, San Francisco.

Wallace, C.S., and D.M. Boulton. 1968. An information measure for classification. *Computer Journal* 11:185–194.

Zimmer, Karl E. 1969. Psychological correlates of some Turkish morpheme structure conditions. *Language* 45:309–321.

Zuidema, Willem. 2003. How the poverty of the stimulus solves the poverty of the stimulus. In *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)*, ed. Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, 51–58.