

On evaluation metrics in Optimality Theory

Ezer Rasin and Roni Katzir
rasin@mit.edu, rkatzir@post.tau.ac.il

September 16, 2013

1 Background

A central component of Chomsky and Halle (1968)'s Sound Pattern of English (SPE) is the evaluation metric, a criterion for comparing grammars given the data. For a grammar G that can parse the data, the value of G is the inverse of the length of G , $\frac{1}{|G|}$. Using this criterion, the child can try to search through the space of possible grammars, eliminating suboptimal grammars as it proceeds.

The evaluation metric was attractive for a variety of reasons. In particular, it is a very general criterion for comparing hypotheses, and it is one that works directly with the representations provided by UG. Different SPE-style theories of UG might be entertained by the phonologist, and these theories may differ in the representations they allow – for example, one theory may sanction braces as a mechanism for abbreviating rules, while another theory will bar such mechanisms – but whichever theory is chosen, the evaluation metric will be able to work with it without modification. This, in turn, opens the way for the phonologist to use the evaluation metric to compare theories of UG: two proposals might be comparably adequate in accounting for adult data, but the predictions regarding learning, given the evaluation metric, may well diverge.

Despite its conceptual appeal, the evaluation metric did not lead to actual learning algorithms for SPE. In part, this can be attributed to specific choices in the definition of the metric, chief among them the decision to treat all grammars that parse the input as equally successful empirically, leaving the evaluation entirely to the prior preference for simpler grammars. As we will see, this choice leads to serious deficiencies with respect to choosing between hypotheses. A growing discomfort within generative linguistics regarding infinite learning spaces may have also played a role. In the years following SPE, the evaluation metric was quietly abandoned.

The advent of Optimality Theory (OT; Prince and Smolensky, 1993) has led to an explosion of work on learning in phonology, including several concrete learning algorithms. For the most part, however, this work has targeted specific aspects of the learning problem and has not been carried out within the framework of an overarching evaluation metric. This is not surprising: the original formulation of the evaluation metric makes it straightforward to apply to rule-based systems but much less easy to apply to constraint-based ones.

Our goal in this paper is to develop an evaluation metric for OT. The conceptual advantages of such a metric are as significant for OT as they were for SPE. In particular, the generality of the evaluation metric holds the promise of working across components, encompassing the lexicon, the constraints, and morphology, among other parts of the grammar. And the close connection to the representations makes the evaluation metric a natural starting point for the study of learning: in a sense, this metric will be what we are entitled to simply by virtue of committing to the kind of UG specified for by OT.

Our empirical focus in this paper is the lexicon and the constraints. We wish to model aspects of knowledge such as the English-speaking child's knowledge that *k^haet* is underlyingly *kæt*, that *raiDer* is underlyingly *raiter*, and that *rai:Der* is underlyingly *raider*. We take it that any theory of phonology would require this knowledge to be learned rather than innate, making this a convenient place to start. The learner we will present succeeds in obtaining this kind of knowledge, which, to our knowledge, makes it a first.

The generality of the evaluation metric will allow us to learn additional parts of the grammar without changing our learner. We will demonstrate this by learning not just the lexicon and the ranking of the constraints but also the constraints themselves. Here not all theories agree that this knowledge is learned – indeed, classical OT assumes that the content of the constraints is innate. However, recent work by Heinz (2007) and Hayes and Wilson (2008) has shown that the acquisition of phonotactic knowledge is a rich and interesting question, and we believe that learning the content of the OT constraints (both markedness and faithfulness constraints) from general constraint schemata is at the very least a direction worth exploring. The learner that we present succeeds in obtaining this knowledge, making it a first in this domain as well.

We start, in section 2, by constructing the evaluation metric, using the analogy of the child as a scientist. By noting the steps that a phonologist would go through in analyzing an unknown language we will obtain a recipe for the simultaneous induction of lexicon, constraints, and ranking. We will point out that the different steps of the recipe can be unified by observing that they all involve the optimization of two quantities, one that reflects the compactness of the grammar itself (including the lexicon) and one that reflects the ease with which the grammar can be used to describe the data. The evaluation metric for the phonologist will be the sum of the two quantities.

After developing the metric for the phonologist and noting its close connection with a line of work pioneered by Solomonoff (1964a), we will suggest that this metric can form the basis for an evaluation metric for the learner. The phonologist and the learner are different in many significant ways – for example, the child's hypothesis space is constrained by UG, while the phonologist is free to explore arbitrary hypotheses; and the phonologist may perform experiments, which the child cannot – but we will argue that the same criterion that defines the evaluation metric within the space of hypotheses entertained by the phonologist can guide the learner within the space of hypotheses defined by UG. Moreover, we will argue that, for the linguist studying learning, this should be the null hypothesis regarding the child.

In section 3 we will present preliminary simulation results. We will try to show, using four different datasets generated by artificial grammars, that the evaluation metric enables the successful learning of nontrivial combinations of lexicons and constraints.

In section 4 we review previous proposals for learning within OT. As we discuss in section 4.1, most of the work in the literature focuses on questions that are quite distinct from those of the present paper. We then turn to two approaches that are much closer to our own in their aims: Maximum-Likelihood Learning of Lexicons and Grammars (Jarosz, 2006b,a), section 4.2; and Lexical-Entropy Learning (Riggle, 2006), section 4.3. We will show that these proposals can be understood in terms of the evaluation metric and that this move highlights inherent difficulties for each of the proposals. Section 5 concludes.

2 The proposal

2.1 *ab-nese*

Consider a phonologist faced with the task of analyzing a newly discovered language. Suppose that the phonologist is working with an informant, who produces the following strings:

- (1) *bab, aabab, ab, baab, babaaaa, bababababababab, aaab, babababaa, babaaaa, aaab, bababababababab, baab, bab, ab, aabab, aabab, baab, bababababababab, aaab, babababaa, ab, babaaaa, bab, aaab, ab, aaab, aabab, babababaa, baab*

Ahead of examining the data in (1), the phonologist might take an uncommitted stance according to which any sequence of humanly pronounceable segments is equally plausible. After a quick glance at the data, however, the phonologist is struck by the following observation: of all the phonetically realizable segments, only *a*'s and *b*'s appear in the strings produced by the informant. This can be seen as an overgeneration problem for the preliminary, uncommitted hypothesis: in the absence of anything within the grammar to rule out the appearance of segments such as *c*, *d*, and *e*, their absence from (1) has to be treated as a surprising accident. The phonologist concludes that this absence is not an accident and that the new language, call it *ab-nese*, prohibits any segment other than *a* or *b*. Within the framework of OT, this restriction can be expressed by positing markedness constraints of the form $*c$, $*d$, $*e$, etc., which we will abbreviate as follows:

- (2) Constraints: $*\neg\{a, b\}$

The phonologist may wish to support (2) by running experiments of various sorts. For example, the phonologist may confront the speaker with two novel forms, one composed only of *a*'s and *b*'s and the other including some other segment as well. To keep the discussion simple, let us assume that if the phonologist runs such experiments then, both here and in what follows, the results support the generalizations made so far.

The constraints in (2) correctly rule out any string that includes segments other than *a* or *b*, thus solving the potential overgeneration problem. As it stands, however, the analysis in (2) still suffers from an overgeneration problem: in the sequence in (1), certain sequences of *a*'s and *b*'s, such as *ab* and *babababaa*, appear multiple times, while other sequences of *a*'s and *b*'s, such as *baba* and *abb*, never appear at all. The strings that repeat themselves are the following:

- (3) 1) ab 3) aaab 5) baab 7) babababaa
 2) bab 4) aabab 6) babaaaa 8) babababababaabab

To remedy this overgeneration problem, the phonologist conjectures that the grammar of *ab*-nese includes a *lexicon*, a repository for information about the specific forms that are allowed. As a simple starting point, the phonologist posits (3) as the lexicon. Within the framework of OT, restricting the grammar to forms generated from a lexicon does not immediately address the overgeneration problem: selections from the lexicon can, in principle at least, surface as any form; a single entry in the lexicon can thus generate any conceivable output. In order to ensure that this does not happen and that the elements selected from the lexicon surface unchanged, the phonologist also posits a constraint, *FAITH*, that penalizes a surface form that is not identical to the underlying form, taken from the lexicon, that was used to derive that surface form:¹

- (4) Constraints: $*\neg\{a, b\}$, *FAITH*

Given the lexicon in (3), the constraints in (4) are unviolated, and so no ranking among them is needed at this point.

With the aid of (4), the phonologist now has a grammar that accounts for the fact that the data in (1) are instances of the entries in (3). The analysis is not fully satisfactory, however: it misses what seems like a significant generalization, namely that two *b*'s never appear in a row. The phonologist characterizes the generalization in terms of an additional markedness constraint, **bb*, which is ranked above *FAITH* to ensure that *bb* sequences in the lexicon will not survive the mapping to surface forms:

- (5) Constraints: $*\neg\{a, b\}$, **bb* \gg *FAITH*

The phonologist now sets out to change the lexicon so as to take advantage of the new markedness constraint **bb*. There are two different ways in which **bb* could generate one of the surface forms in (3) from an underlying form that violates the constraint. When two occurrences of *b* are adjacent in the underlying form, **bb* could cause one of them to delete, or it could cause an occurrence of *a* to be epenthesized between them. Taking the surface form *aabab* as an example, the two possibilities are the following:²

- (6) 1) *a* epenthesis $bb \rightarrow bab$ $/aabb/$
 2) *b* deletion $bb \rightarrow b$ $\{ /aab^n ab^m / : n, m \geq 1 \}$

¹When we introduced $*\neg\{a, b\}$ in (2) above, its purpose was to address the initial overgeneration problem that we encountered. Now, with the introduction of the lexicon and of an undominated *FAITH*, this problem is resolved independently of $*\neg\{a, b\}$. This does not mean that $*\neg\{a, b\}$ has become redundant, however: if the phonologist fails to take the restriction on the segmental inventory in *ab*-nese into account, the fact that the lexicon is written only in *a*'s and *b*'s will have to be taken to be a surprising accident; with the commitment to $*\neg\{a, b\}$, on the other hand, the lexicon seems much more natural. In other words, the present step involves a subtle but significant shift in the role of $*\neg\{a, b\}$ from an aid in making the raw data look more natural to an aid in making the lexicon look more natural. This shift raises interesting issues regarding the architecture of the lexicon, but we hope that further discussion of these issues and their resolution can wait for a separate occasion.



²The deletion option in (6) is shorthand for an infinite family of possibilities for the lexicon.

Deletion and epenthesis are both possible in principle (indeed, it is possible for both to be active within the same grammar), but the reasonable phonologist will presumably take epenthesis as the default analysis for all the cases above, switching to a deletion analysis only in case of additional evidence that such a move is needed. Informally, this preference for epenthesis in this case follows from considerations of economy: on the assumption that *a* is epenthesized between two adjacent *b*'s, the lexicon is smaller than it is on the assumption that it contains additional *b*'s that are deleted. The underlying forms, then, are as follows:

- (7) 1) /ab/ 3) /aaab/ 5) /baab/ 7) /bbbbaa/
 2) /bb/ 4) /aabb/ 6) /bbaaaa/ 8) /bbbbbaabb/

The lexicon in (7) is a clear improvement over (3): an intuitively significant regularity, namely the absence of two consecutive *b*'s, has now been squeezed out of the lexicon, supporting a more economical representation. Note, however, that the constraints in (5) do not allow us to take full advantage of the improved lexicon. The ranking of **bb* over FAITH allows us to correctly generate all of the observed forms, using *a*-epenthesis where needed, but it also allows us to employ *b*-deletion and map URs including the forbidden sequence *bb* onto other, unattested forms. For example, the UR *bb* can be mapped either to the attested *bab* (through epenthesis) or to the unattested *b* (by deletion).

(8)

	/bb/	$*\neg\{a, b\}$	<i>*bb</i>	FAITH
a.	bb		*!	
b. 	b			*
c. 	bab			*

In other words, by economizing the lexicon we have introduced a new overgeneration problem.

Fortunately, the new overgeneration problem can be resolved at the cost of a very minimal further complication of the grammar. To ensure that only *a*-epenthesis resolves double-*b* sequences, we can split FAITH into two faithfulness constraints: MAX, which penalizes deletions; and DEP, which penalizes insertions.³ We can now rank **bb* above DEP but not above MAX, ensuring that avoiding *bb* will justify insertion (of *a*) but not deletion (of *b*):

- (9) Constraints: $*\neg\{a, b\}$, MAX, **bb* \gg DEP

Is the analysis complete? The answer is yes, but it will be useful to understand why. The steps we took in developing the analysis above were meant to address two kinds of concerns: we wanted to minimize overgeneration with respect to the attested forms; and we wanted to avoid any pointless complexity in the analysis itself. Let us call the first consideration *restrictiveness* and the second *economy*. As far as the data in (1) are concerned, the analysis, combining the lexicon in (7) and the constraints in (9), seems fully restrictive: it can generate only those forms that have been observed. It is always conceivable, of course, that future observations will force a modification of

³To simplify the present discussion, we consider here only insertions and deletions as possible modifications.


the analysis, but this is no more than the usual state of affairs in an empirical science. Restrictiveness, then, seems to be satisfied. What about economy? We just saw that **bb* allowed us to obtain a more compact theory. But this was just one among many patterns in the data, and it might seem tempting to try to capture some of the additional patterns as well. For example, the astute reader will have noticed that the number of *b*'s in the examples is always a power of 2: 1, 2, 4, and 8 (higher powers are missing). And the number of *a*'s is always a Fibonacci number: 1, 2, 3, 5, and 8 (all Fibonacci numbers higher than 8 are missing). Somewhat less exotically, the sequence *aaaaa* never appears in the data, and the sequence *ba* never appears word-finally. In principle at least, there is nothing to prevent us from modifying the grammar so as to take advantage of these patterns and squeeze them out of the lexicon.⁴

At this point, however, capturing any of these patterns will probably be more trouble than it is worth. If we keep discovering new words in *ab-nese*, and if these words conform to one of these patterns, it will eventually become important to capture it. But for now, these facts, and infinitely many additional ones, do not help make the analysis simpler and are best treated as accidents rather than meaningful patterns: the analysis, as far as the current data are concerned, is complete.

2.2 Formalizing the process: an evaluation metric for the phonologist

The process just described attempts to maximize the economy and the restrictiveness of the grammar given the data. In section 2.3 we will use the phonologist's criterion for comparing hypotheses – the phonologist's evaluation metric – as a model for the learner's evaluation metric. Before we can do that, however, we will need to make the phonologist's evaluation metric explicit and precise. In particular, we will need to understand how economy and restrictiveness are measured and how the two measurements are combined. As it turns out, it is easy to make incorrect choices here, choices that would lead the phonologist to favor hypotheses that clash directly with our intuitions regarding linguistic analysis. We will see a few illustrative cases below. But let us start with what we think is the right choice, first formulated by Ray Solomonoff (Solomonoff, 1960, 1964a,b). According to Solomonoff, a hypothesis is a complete description of the data – think of it as a computer program that runs, prints out the

⁴For example, we could add the following markedness constraints to the grammar and rank all of them above DEP: FIB(a) (penalizing any form in which the number of *a*'s is not a Fibonacci number), 2ⁿ(b) (penalizing any form in which the number of *b*'s is not a power of 2), **ba#*, and **aaaaa*. We can use these constraints to obtain a shorter UR for the surface form *aabab*.

	/aaba/	*¬{a, b}	MAX	*bb	FIB(a)	2 ⁿ (b)	*ba#	*aaaaa	DEP
	a. aaba						*!		
i.	b. aab		*!						
	c. aabaa				*!				*
	d.  aabab								*

In order to save a single segment in this UR, we needed two new constraints – FIB(a) and **ba#* (the remaining two new constraints were not involved in this case). This is hardly a bargain deal, and it does not improve much through consideration of the remaining forms in the lexicon. If enough additional forms of the general pattern exhibited by *aabab* are encountered, however, the resulting savings in the storage of the URs will merit the price paid by introducing the new constraints.

data, and then halts. The value of a hypothesis is determined by its length: the shorter the hypothesis (for example, as measured in bits in the source file containing it on the computer), the better it is. It is often convenient to separate the logic of the program from any accidental aspects of the data and think of the program as the combination of two distinct parts: the ‘real’ program, or grammar, which we will write as G ; and the encoding of the data D using the grammar, which we will write as $D|G$. As we will shortly see, the length of G , $|G|$, corresponds to the informal notion of economy, while the length of $D|G$, $|D|G|$, corresponds to restrictiveness. The goal of the phonologist, on this view, is to find the hypothesis that provides the shortest overall length. That is, the grammar that provides the shortest value for the sum $|G| + |D|G|$.

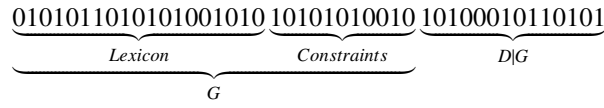


Figure 1: Schematic view of Solomonoff’s evaluation metric as applied to OT. The grammar G consists of both lexicon and constraints. The data D are represented not directly but as encoded by G . The overall description of the data is the combination of G and $D|G$.

Let us illustrate. Suppose we wish to obtain a complete description of the data in (1), for example in order to convey it to a phonologist who has no direct access to our informant. Ahead of the analysis that we went through, the data would be no more than an arbitrary sequence to us. To convey it we would be able to do no better than transmit it symbol by symbol, specifying at each step which symbol is chosen out of the full alphabet. The usual way of specifying choices out of a set is as a string of bits – that is, a string of binary choices, each of which can be 0 or 1. If the full alphabet has four elements, for example, we can arrange them in a row – say, a_1 , a_2 , a_3 , and a_4 – and specify the choice using two bits: the first specifying whether the choice is among the leftmost two or the rightmost two (so 0 says that the choice is among a_1 and a_2 and 1 says that the choice is among a_3 and a_4) and the second doing the same within the subset specified by the first (so if the first bit was 0 and the second bit was 1, then the specified element is a_2). If there are eight elements in our full alphabet, write them as a_1 to a_8 , two bits would no longer suffice: we would need an additional bit to specify first whether the chosen element is among the leftmost four or the rightmost four, after which two bits will allow us to specify the exact choice as before. More generally, if there are n elements in our full alphabet, we would need $\lceil \lg n \rceil$ bits to specify an individual element. For example, if our alphabet is the IPA, which has 107 letters and 31 diacritics, we would need eight bits to encode an individual choice. To convey the data in (1) under the null hypothesis, then, we would need to spend the number of bits we require to encode an arbitrary symbol – eight if we are using the IPA – times the number of characters in the sequence, including commas.

As soon as we notice that only a ’s, b ’s, and commas occur in the input sequence, we can replace the eight bits per symbol with a fixed code length of two bits per symbol,

and the length of the code drops accordingly.⁵ Encoding the restriction of the segmental inventory to the set $\{a, b\}$ takes up a few additional bits, but this addition is easily offset by the savings obtained through the drop from eight bits to two, even for a relatively short text. The comparison between the two hypotheses is schematized in Figure 2.

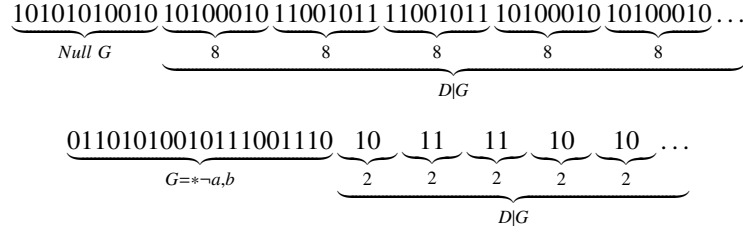


Figure 2: Two simple hypotheses (schematic). The null hypothesis (top) treats the data as an arbitrary sequence of segments. Encoding the grammar is simple, but the price paid for encoding the data is high: eight bits per segment. The hypothesis that treats the data as an arbitrary sequence of a 's, b 's, and comas requires a slightly more complex grammar, but the savings in encoding the data are noticeable: we now have to pay only two bits per segment.

Our next step in the analysis, introducing a lexicon, allows us to derive further savings. If there are only eight sequences that keep repeating themselves, we no longer need to encode each segment individually. Instead, we can transmit the lexicon once, in the beginning of the transmission, and then use $\lg 8 = 3$ bits to specify which word is chosen each time. For *babaaaa*, for example, this would mean three bits instead of fourteen bits for each occurrence.

Observing that sequences of the form *bb* are systematically absent allows us to compress the lexicon introduced in the previous step: we increase the size of the grammar slightly, by adding the constraint **bb*, and this allows us to decrease the overall size of the grammar by removing inter-*b* instances of *a*. (Note that this trade-off is carried out entirely at the level of economy; we will immediately turn to the effect of this move on restrictiveness.) Here the savings are not as dramatic as they were in the previous steps, though they might still be meaningful, and they would be even more so with a bigger lexicon (assuming it conformed to the same pattern).

Next, as long as FAITH is kept as an atomic constraint, we face an overgeneration problem that would leave us worse off than with the uncompressed lexicon. Each time the UR *bb* is selected in order to produce the surface form *bab*, the system so far would generate two winning candidates, the attested *bab* and the unattested *b*. We would thus have to spend additional bits to ensure that we produce the former and not the latter. We overcome this problem by splitting FAITH into two separate constraints, MAX and DEP, and by ranking **bb* above the latter but not above the former. The splitting of FAITH slightly increases the size of the grammar, but it is a one-time increase, and after that every time the UR *bb* is selected, it will lead deterministically to the surface form *bab*. The past three steps are schematized in Figure 3.

⁵We ignore here the slight additional savings made possible by using a variable code length.

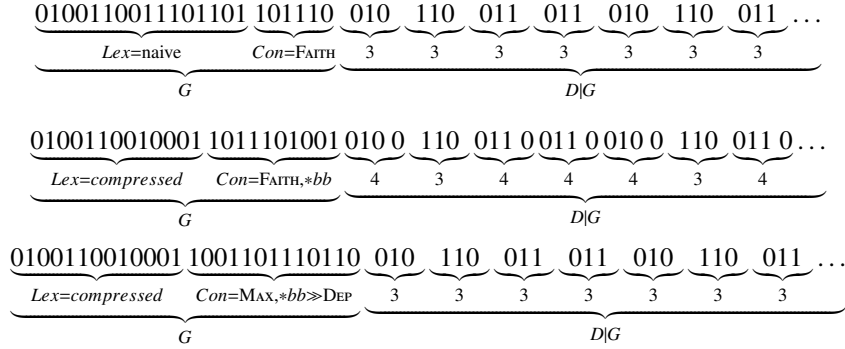


Figure 3: Three more advanced hypotheses. Introducing a naive lexicon, in which the attested strings are listed, allows us to describe the data word-by-word rather than segment-by-segment, yielding significant savings (top). Squeezing the pattern **bb* out of the lexicon results in a shorter lexicon but longer overall description length: for each UR that includes the sequence *bb*, we will now need to specify that the surface form is the result of *a*-epenthesis rather than *b*-deletion (middle). Splitting *FAITH* into *MAX* and *DEP* allows us to maintain both a short lexicon and a short description of the data at the modest cost of a slight complication of the constraints, leading to the shortest overall length (bottom).

Finally, the putative patterns of powers of 2 and the Fibonacci sequence seem quite unhelpful at this point in terms of compression, as do **a⁵* and **ba#*. Differently from **bb*, which aided in compression and was thus taken to capture a meaningful gap, these other patterns are taken to capture accidental gaps and are thus not added to the grammar.

Each of our steps above attempted to improve the analysis by shortening the encoding. In this respect, the phonologist's strategy is one among many imaginable strategies incorporating a simplicity bias, a general approach that is often associated with Occam's Razor. But details matter. Crucially, what matters to the phonologist is the *entire* message length: economy (that is, the length of the grammar, including the lexicon) and restrictiveness (that is, the encoding of the data given the grammar) must be balanced against one another. Minimizing only the one or only the other would lead to unhappy results. Suppose, for example, that we had focused on economy alone. We would then never have departed from the initial, perfectly simple hypothesis that said that any sequence of segments is possible. And if we had been forced by someone to abandon that hypothesis and accept that only *a*'s and *b*'s occurred, we would have settled on that hypothesis and moved no further. If we had been forced to move forward and adopt a lexicon, we might have had an incentive to minimize it by adding **bb* to the grammar and shortening the URs, but we would have had no cause to split *FAITH* into *MAX* and *DEP*. Economy alone, then, is insufficient. Note that combining a first step of economy with a second step of restrictiveness will be of little help: the problematic winner in each of the steps just summarized is strictly simpler than the losing competitor, thus making it impossible for some tie-breaking criterion in the second step to reverse the unfortunate outcome. Economy alone is the essence of the evaluation metric of Chomsky and Halle (1968), and a two-step architecture in which a criterion

such as restrictiveness operates on the outcome of economy is at the heart of the earlier version of the evaluation metric in Chomsky (1951), as well as what Kiparsky (2007) calls Pāṇini’s Razor. The problem for economy has been noted by Braine (1971), Baker (1979), and Dell (1981) and was among the reasons for the retreat from the evaluation metric in generative grammar.

Consider next what would happen if the learner chose to ignore economy and focus on restrictiveness alone. As we just saw, economy alone gives the learner no incentive to leave a simple but incorrect hypothesis that is compatible with the data. This occurs due to the fact that simple grammars are often very inclusive, admitting a proper superset of the language at question. This, in turn, means that the language in question cannot provide counterexamples to the learner’s incorrect hypothesis. This is an instance of the so-called *subset problem*, and it is a problem that learners based on economy alone are vulnerable to. In order to counter the dangers of the subset problem, learners often incorporate a preference for conservative guesses. In particular, it is sometimes suggested that learners should always choose the smallest language compatible with the data, a preference known as the *subset principle*. Restrictiveness alone is an approach that respects the subset principle. While escaping the subset problem, such a learner runs straight into the mirror-image of the problem for economy alone: instead of wild overgeneralization, such a learner never generalizes at all. In the case of *ab-nese*, for example, we would have been perfectly content with our first lexicon, and nothing would have made us add **bb* and compress the URs. This, in turn, would make a putative future word *aab* equally easy to accommodate as *abb*; and while *ab-nese* is of course artificial, the counterparts of this prediction for natural languages such as English have been recognized as problematic as early as Halle (1962). The dangers of adhering to the subset principle become particularly clear when the language is infinite (or just too big for the phonologist to encounter in its entirety). To keep things simple, imagine a dialect of *ab-nese*, call it *zab-nese*, in which any nonnegative number of *z*’s can precede any word. We would expect a reasonable phonologist to notice this generalization after enough input elements have been observed. A fully restrictive phonologist, however, will never generalize. At any given point, such a phonologist will have had exposure to a finite number of such *z*-variants, and these forms will be listed. Note also that, as with our earlier discussion of economy, the problem will not be solved by using restrictiveness as a first step that then feeds a second criterion such as simplicity. The incorrect winner at each step in the case of *zab-nese* will always be strictly more restrictive than the correct hypothesis, rendering a second step useless.

In short, we must take both economy and restrictiveness into account, and we must minimize both simultaneously: our goal is to minimize $|G| + |D|G|$. As mentioned, the first to propose this idea was Solomonoff (1964a), who used his discovery to formulate a fully general theory of prediction. The same idea of viewing hypotheses as programs that output the data and defining their value according to their length was discovered independently (from a slightly different perspective) by Kolmogorov (1965) and Chaitin (1966). The length of the shortest program that outputs the data *D* is known as the *Kolmogorov complexity* of *D* and is written $K(D)$.⁶ Kolmogorov complexity is not computable, and it is often necessary to restrict the hypothesis space

⁶See Li and Vitányi (2008) for a detailed and thorough discussion of Kolmogorov complexity.

to ensure computability. This is done in the frameworks of Minimum Message Length (MML; Wallace and Boulton, 1968) and Minimum Description Length (MDL; Rissanen, 1978). To simplify terminology, and since the differences between the frameworks incorporating Solomonoff’s insight will not be central to our proposal, we will refer to any attempt to minimize $|G| + |D|G|$ (often within a restricted family of possible grammars) as MDL. The relevance of MDL for grammar induction was already noted by Solomonoff (1964b). Over the years, several authors have used MDL profitably for grammar induction, either as a methodological principle for the scientist or as a learning criterion for the learner. Notable examples include Berwick (1982), Rissanen and Ristad (1994), Stolcke (1994), de Marcken (1996), Grünwald (1996), Clark (2001), Goldsmith (2001, 2010), Dowman (2007), Chater and Vitányi (2007), and Hsu and Chater (2010).

2.3 From phonologist to learner

A learner is not a phonologist. The phonologist may, in principle at least, consider any program as a grammar; the learner, on the other hand, may well be restricted by UG to a very limited search space. Earlier, for example, we suggested that the phonologist may consider – but ultimately reject – four patterns in the *ab*-nese data: the number of *a*’s is always a Fibonacci number; the number of *b*’s is always a power of two; the sequence *aaaaa* never occurs; and the sequence *ba* never occurs word-finally. The phonologist, being a scientist, may consider all four patterns. And while it is conceivable that the learner is also capable of entertaining all these patterns, it is also possible that some of these patterns are impossible for the learner to entertain and thus lie outside its search space. Another difference between the phonologist and the learner is the degree of control each has on their respective inputs: as mentioned earlier, the phonologist may run controlled experiments using a variety of methodologies, recruit typological data, and obtain systematic negative evidence; the learner, on the other hand, is largely restricted to the kinds of evidence that are given to it by its environment. An additional difference is that the child might be restricted to use a particular search strategy in traversing the hypothesis space, while the phonologist is free to search in any way they wish. And the phonologist may also record many years of data and make reference to all the information accumulated in this fashion, while the child is quite unlikely to record explicitly the entire history of speech to which it has been exposed.⁷

But in a certain sense, the learner and the phonologist have a great deal in common. Both face the task of making sense of unanalyzed data in the language they are immersed in, and both bring to the task a hypothesis space, each point in which representing a mechanism – equivalently, a grammar – for generating the data. Not all the mechanisms in the hypothesis space will be able to generate the data in the first place, but for any mechanism that can, we can look for a sequence of instructions to the mechanism – a key – that will generate exactly the part of the data that we have seen. As discussed above, a message consisting of the combination of a grammar and a key

⁷The differences between the learner and an ideal scientist are the focus of the growing literature on underlearning, a literature that investigates the limitations of what humans can learn. See Smith (1966), Peña et al. (2002), Endress et al. (2007), Moreton (2008), Endress et al. (2009), Endress and Mehler (2010), and Becker et al. (2011), among others.

provides a full description of the data, and we can think of the phonologist as searching the space for the shortest such message. The hypothesis space for the phonologist is biased toward mechanisms that work well with past observations – recall that in discussing *ab-nese* we took it for granted that we could easily encode constraints, ranking, and lexicons – but it is a very big space, and it includes many additional mechanisms (for example, if *ab-nese* turned out to be problematic for OT, we might consider a complete revision of the architectural premises of the grammar). This was the essence of the discovery procedure that we built in the previous section.

For the child learner, things are less clear. Like the phonologist, the child attempts to settle on a point in the hypothesis space (in the case of the child, the hypothesis space is probably considerably more limited than for the phonologist; for example, either OT or SPE but presumably not both). But there is little conclusive evidence to date about how the child chooses this point. It is conceivable, of course, that the child searches through its hypothesis space in exactly the same way that the phonologist searches through theirs. But there are any number of other procedures that the child might be using. For example, it might look for the first grammar under some enumeration that is compatible with the data; or it might look for a grammar that is *not* compatible with the data; or it might use the first two words in the input as a key for selecting a grammar out of a big table, regardless of compatibility; and so on. Of course, there are also learning procedures that are considerably more reasonable than those just mentioned. See, in particular, Manzini and Wexler (1987), Gibson and Wexler (1994), and Niyogi and Berwick (1996) for such procedures for the hypothesis space provided by the Principles and Parameters framework of Chomsky (1981); we will shortly review several procedures of this kind that have been proposed for the hypothesis space provided by OT.

However, and this is a point we wish to stress, of all the different procedures the child might be using, the one that parallels the phonologist's search, when applied to the hypothesis space provided by UG, is in a sense the simplest. The child already has access to the hypothesis space, and each point in that space allows it to parse the data; all that is missing is the ability to traverse this space and test different hypotheses, comparing the messages they support in terms of overall length. If the child can maintain a current hypothesis and a new hypothesis simultaneously and use them both to parse the data, and if the child can switch from one hypothesis to another in a way that lets it traverse a portion of the hypothesis space that allows convergence, it will be able to mirror the phonologist's search. And if the child can compare the overall memory space required to encode the data using two hypotheses, it can mirror the phonologist's criterion. The procedure that parallels the phonologist, then, is available to the child almost in full simply by virtue of having the ability to represent and use grammars from within the set allowed by UG: indeed, it seems that one would have to make special stipulations to block such a procedure. Any other procedure for selecting grammars that we know of requires considerably bigger commitments. As a matter of scientific methodology, then, it makes sense to take the child-as-phonologist model as the null hypothesis and abandon it in favor of other, more elaborate models only in the face of sufficient evidence.

3 Simulation results

One obvious kind of evidence for the inadequacy of the child-as-phonologist model would be a demonstration that it is incapable of learning the kinds of patterns that children acquire from typical data, and that some other learning procedure manages to learn these patterns better. Our next step, then, will be to examine the behavior of a learner that implements the child-as-phonologist model on several different datasets. We will not be able to test the learner on a real-life corpus at this point.⁸ Instead, we will provide a proof-of-concept demonstration, using datasets generated by artificial grammars that incorporate phonological dependencies that we consider interesting. We will start with the *ab-nese* dataset, move on to a language that exhibits some phonological patterns familiar from English, continue to a dataset showing restricted optionality, and end with a dataset exemplifying the ability of the current approach to learn from alternations. As we will see, the learner extracts the correct grammars in all four cases.

We need to commit in advance to the search space defined by UG: here we will assume that this space is defined by the ability to state lexicons using a fixed alphabet of feature vectors and the ability to state constraints (and their ranking) using two kinds of very general constraint schemata, one for faithfulness constraints and one for markedness constraints, as shown in Figure 4.⁹ We wish to emphasize, though, that our goal is not to argue for this particular theory of UG but rather to demonstrate how learning can take place given a search space provided by UG and an MDL evaluation metric.

$$\text{DEP}(F) \quad \text{MAX}(F) \quad \text{IDENT}(F) \quad *F_1 F_2 \dots F_n$$

Figure 4: Constraint schemata available to the learner. F 's represent feature bundles.

Our focus in this paper is the learning criterion. We have nothing to say about either the search procedure or the initial state of the search. To make the learner concrete, though, we must make commitments with respect to both. For the search procedure, we adopt Simulated Annealing (Kirkpatrick et al., 1983). For the initial state, we assume the naive one in which no patterns in the data have been discovered. The grammar includes a single faithfulness constraint *Faith* that penalizes all structural changes, thus enforcing an identity mapping between URs and surface forms.¹⁰

⁸We will not attempt to speculate on the amount of data that the child may refer to (with one extreme being an unbounded batch learner, the other a memory-less online learner, and real life presumably somewhere in between). The learner presented here is a batch learner, but the amount of memory that it uses for the data in the following examples is relatively small. We hope that an investigation of the amount of data used by the human learner and of whether the current learner can be modified to match this memory constraint can wait for a separate occasion.

⁹The fixed alphabet could be part of the innate endowment of the learner. Alternatively, it could be learned during an earlier phase of learning. As far as the present discussion is concerned, all that matters is that the alphabet is fixed.

¹⁰In the literature following Smolensky (1996), an initial ordering of Markedness over Faithfulness ($M \gg F$) is often assumed as a means to confront the subset problem, but see Hale and Reiss (1998) for arguments in favor of a faithful initial state. See Albright and Hayes (2011) for further relevant discussion. On the

3.1 *ab-nese*

Our first dataset is a language similar to *ab-nese*, presented in section 2.1 above and repeated here:

- (10) *bab, aabab, ab, baab, babaaaa, bababababababab, aaab, babababaa, babaaaa, aaab, bababababababab, baab, bab, ab, aabab, aabab, baab, bababababababab, aaab, babababaa, ab, babaaaa, bab, aaab, ab, aaab, aabab, babababaa, baab*

Given an alphabet $\Sigma = \{a, b\}$ and one feature $\pm cons$ ($a = [-cons]$, $b = [+cons]$), we generated a few hundred words by taking combinations of segments in Σ . We then filtered out all words that included the sequence *bb* and provided the learner with the resulting set of words. The initial state comprised a constraint set with a single FAITH constraint and a lexicon identical to the data:¹¹

- (11) Initial grammar:

$$G_{initial} = \begin{cases} \text{LEX:} & \textit{bab, aabab, ab, baab, babaaa, babababaa} \dots \\ \text{CON:} & \text{FAITH} \end{cases}$$

$$\text{Description length: } |G_{initial}| + |D|G_{initial}| = 9,230 + 403,200 = 412,430$$

As discussed in section 2.1, the absence of *bb* sequences from the data can be used to obtain a more concise description of it. Consequently, the evaluation metric favors grammars that encode this pattern over grammars that treat it as a mere accident. Our learner converged on a final hypothesis that had relevant instances of *a* removed from the lexicon and inserted by the grammar:

- (12) Final grammar:

$$G_{final} = \begin{cases} \text{LEX:} & \textit{bb, aabb, ab, baab, bbaaa, bbbbaa,} \dots \\ \text{CON:} & \text{MAX}([+cons]) \gg * [+cons] [+cons] \gg \text{FAITH} \end{cases}$$

$$\text{Description length: } |G_{final}| + |D|G_{final}| = 4,010 + 403,200 = 407,210$$

The addition of both the markedness ($* [+cons] [+cons]$) and the faithfulness ($\text{MAX}([+cons])$) constraints increases the length of CON but helps in minimizing the total description

current proposal, restrictiveness is obtained as a by-product of the MDL evaluation metric rather than as a property of the initial state.

¹¹We limited the words to be up to six syllables long, though nothing hinges on this choice. We also chose to multiply the summand $|D|G|$ by 100 in these simulations due to performance considerations. The number of bits required to describe the data given the grammar is affected by the amount of data the learner is exposed to. By multiplying this factor by a large number we avoided working with very large corpora that would have significantly increased the running time of our algorithm. From a computational complexity point of view, the problem that we are dealing here with is NP-hard (see Heinz et al., 2009) and the constants involved are considerable. We believe that the question of whether $|D|G|$ is indeed multiplied by a constant factor is an interesting question that should be empirically investigated. Currently, however, we have nothing substantial to say about this matter.

length. The markedness constraint allows the learner to compress the lexicon by preventing *bb* sequences from surfacing. The faithfulness constraint is introduced to ensure that *b* deletion incurs more violations than *a* epenthesis. The latter option is therefore deterministically chosen for satisfying the markedness constraint, and the length of the data given the grammar becomes lower than it would have been had the faithfulness constraint been left out. The learner has converged on the correct grammar.¹²

3.2 Aspiration

Our next dataset shows a pattern modeled after aspiration in English and is designed to test the learner on the problem of allophonic distribution. Simplifying, we assume that the ambient language has aspirated stops (like *t^h* and *k^h*) appearing before vowels but not elsewhere. The distribution of aspiration is thus entirely predictable. We expect the learner to treat aspirated stops as allophones of their unaspirated counterparts. Aspiration should be completely removed from the lexicon and inserted by the grammar in the right context. Thus, the UR of *k^hat* should be */kat/* and the UR of *t^hik^hit* should be */tikit/*, while surface forms where aspiration is missing in the right context (like **kat*) should be ungrammatical. Importantly, the grammar should also block aspiration from occurring elsewhere, as in the illicit surface forms **at^h* and **k^hik^ht*.

Previously, we explained why the MDL evaluation metric favors grammars that treat such patterns systematically rather than leaving them as accidents of the lexicon. Adding the relevant constraints to CON increases its description length but makes it possible to squeeze information out of the lexicon, thereby minimizing the total description length. Here, blocking of aspiration in elsewhere contexts presents a further learnability challenge. The crucial point is that a grammar that generates aspirated stops before vowels is not necessarily restrictive enough; the grammar should also prevent cases where URs like */at^h/* or */kik^ht/* surface with stray aspirated segments.

One way for the learner to approach this problem is to allow forms like *at^h* and *kik^ht* to be represented underlyingly and block **at^h* and **k^hik^ht* as part of the input-output mapping. This direction, in line with the OT principle of Richness of the Base (ROTB), is not available to our MDL learner: on natural assumptions about the representation of aspiration, a hypothesis with additional underlying instances of aspiration will be more complex than one without them and will thus be dispreferred;¹³ and in the absence of such additional instances of aspiration, a constraint that ensures that they do not surface will serve no compressional purpose and will likewise cause the hypothesis to be dispreferred. But constraints on outputs are not the only imaginable response to the restrictiveness problem raised by the aspiration pattern. A different way for a learner

¹²The result differs from the final grammar in our discussion in section 2.1 in two respects. First, a MAX constraint is added instead of having FAITH splitted into MAX and DEP. This occurs since DEP does not require a shorter description length than FAITH and there is no reason for the evaluation metric to favor it. The second difference is that our representations only allow strict ranking of constraints in CON and so MAX([+cons]) can be ranked anywhere in the hierarchy, whereas in our previous discussion we assumed that non-obligatory rankings were possible.

¹³This is true, for example, if aspiration is represented as a separate segment, which is the somewhat simplistic representation we will use below. It is also true on various other, possibly more realistic ways to represent aspiration. It is possible, of course, to choose representations that make it cheaper to encode the presence of aspiration than its absence, but we find it hard to think of a justification for such a choice.

to meet the challenge is to capitalize on the absence of aspiration from the lexicon in order to describe the lexicon more succinctly. If aspiration can be squeezed out of the inventory of primitives from which underlying material is chosen, each choice in the lexicon would cost fewer bits of information. Grammars that ban underlying aspiration will thus rule out URs like $/at^h/$ and $/kik^ht/$ and, consequently, will block surface aspiration in all inappropriate contexts. We can include such grammars in the learner’s search space by enriching the representations and providing the learner with the ability to restrict its inventory of primitives. Such grammars are formally simpler, and if they lead to a lower total description length, they would be favored by our learning criterion. We will now present the learning setting and show that this solution, combined with the MDL evaluation metric, leads to the correct predictions.

The alphabet for our pseudo-English case is $\Sigma = \{a, i, u, p, t, k,^h\}$. We assume that seven features distinguish between the available segments: $\pm cons$, $\pm stop$, $\pm aspirated$, $\pm low$, $\pm round$, $\pm velar$, $\pm labial$. Aspiration is represented as an individual segment $[^h]$.¹⁴ Here, we generated four hundred words varying in length and inserted aspiration after every stop that preceded a vowel. As before, the initial state had one constraint FAITH and a lexicon identical to the data. Note that the segmental inventory is now specified next to the lexicon:

(13) Initial grammar:

a.

$$G_{initial} = \begin{cases} \text{LEX:} & \{a, i, u, p, t, k,^h\}; k^hu, p^hik^ha, t^hipk^hi, p^hiap^hu, kp^hik^hut^ha, \dots \\ \text{CON:} & \text{FAITH} \end{cases}$$

$$\text{Description length: } |G_{initial}| + |D|G_{initial}| = 24,393 + 720,000 = 744,393$$

b. Final grammar:

$$G_{final} = \begin{cases} \text{LEX:} & \{a, i, u, p, t, k\}; ku, pika, tipki, piapu, kpikuta, \dots \\ \text{CON:} & * [+stop] [-cons] \gg \text{FAITH} \gg \text{MAX}([-aspirated]) \end{cases}$$

$$\text{Description length: } |G_{final}| + |D|G_{final}| = 8,904 + 720,000 = 728,904$$

The final grammar includes a markedness constraint that militates against sequences of a stop followed by a vowel ($* [+stop] [-cons]$). Consequently, aspiration is entirely removed from the lexicon and inserted by the grammar in the right context. The segmental inventory has been restricted to $\{a, i, u, p, t, k\}$, blocking aspiration in other contexts as expected; since aspiration cannot be used to describe underlying segments, no UR can derive forms like $*at^h$ and $*k^hik^ht$ (relevant to the examples provided above, aspiration of p in t^hipk^hi and of the first k in $kp^hik^hut^ha$ would be ungrammatical). The allophonic distribution has been correctly learned.

¹⁴We chose to treat aspiration as a distinct segment in order to keep our representations as simple as possible at this stage. This would allow the learner to minimize description length by removing instances of h from the lexicon. Alternatively, the same effect could be achieved by allowing features to be underlyingly underspecified. An unaspirated segment t would not require specification for the feature $\pm aspirated$, making its description length shorter than a corresponding fully specified t^h . In that case, the learner could have altered its feature inventory to squeeze out the feature $aspirated$ instead of squeezing the segment h out of a segmental inventory.

3.3 Optionality

The tension between economy and restrictiveness becomes particularly clear in cases that involve optional phonological processes. The significance of optionality to learnability was articulated by Baker (1979) and Dell (1981), who noted that optionality leads an economy-only evaluation metric, such as that provided in SPE, directly into the subset problem. As mentioned in section 2.2 above, the susceptibility of economy-based learners to the subset problem was chief among the reasons for abandoning the SPE metric. In this section we present a learning simulation modeled after one of Dell’s cases and demonstrate how MDL provides the desired remedy when optionality is concerned.

Let us first consider a concrete example, a modified version of one of Dell’s French examples, before moving on to state the problem more generally.¹⁵ Consider a grammar that handles consonant clusters as follows: an unfavorable sequence C_1C_2 is optionally resolved by either *i* epenthesis between the two consonants or by C_2 deletion. A UR like /*tabl*/ would surface either as [*tabil*] or as [*tab*]. In addition, the grammar generates surface forms that appear as if they could have been derived by the same process, but in fact they are not. For example, the UR /*parl*/ is faithfully mapped into [*paril*], whereas *[*par*] is ungrammatical. A learner exposed to {[*tabil*], [*tab*], [*paril*] } would face the subset problem. On the one hand, it would be justified in making the generalization that [*tabil*] and [*tab*] are generated from the same UR. A learning strategy based solely on economy would succeed in making this inductive leap: a grammar that includes one UR (/ *tabl* /) would be more economical than a grammar that has two URs (/ *tabil* / and / *tab* /), even at the cost of introducing the relevant rule or constraint. On the other hand, if only economy is taken into consideration, a UR like /*parl*/ that is strictly simpler than an alternative /*paril*/ would be preferred. Such a grammar would correctly generate [*paril*] from /*parl*/, but since a consonant cluster could be optionally resolved by deletion, that grammar would also generate the ill-formed *[*par*]. The process involving optionality, which we will refer to as *P*, should not be extended to operating on the UR of [*paril*]. Our target grammar, G_{target} , is strictly simpler than the overly restrictive identity grammar $G_{identity}$, but it has a strictly simpler alternative, call it G_{simple} , that overgeneralizes:

- (14) a. G_{simple} (economy only; overgeneralizing): Admits an overly permissive version of *P*.
- b. G_{target} (economy and restrictiveness balanced; correct): Admits an appropriately restricted version of *P*.
- c. $G_{identity}$ (restrictiveness only; complex grammar; under-generalizing): Does not admit *P*.

The problem faced by the learner, then, is to generalize beyond the data (by applying *P*’s operation to /*tabl*/), but to prevent excessive generalization (by precluding *P*’s

¹⁵We have revised the example to allow an easy formulation of optionality in the OT framework. In OT, optionality could arise when URs have more than one optimal output. Instead of dealing with a process that optionally takes place (but might not apply), we chose to handle a case where a markedness constraint could be resolved by two distinct repairs that are equally harmonic.

operation on the UR of $[paril]$, which would generate the ungrammatical $*[par]$).¹⁶

In terms of MDL, minimizing the size of the grammar would generally be beneficial unless it is counterbalanced by an increased length of data encoding given the grammar. Having to make more choices in the face of optionality results in such an increase. In the case discussed here, the dissimilar grammatical treatment of superficially similar surface forms (*tabil* vs. *paril*) is a consequence of differences in the compression benefits that each one provides. Encoding $[tabil]$ as the output of $/tabl/$ would require paying one bit of information to specify its choice over $[tab]$. Generally, collapsing $[tabil]$ and $[tab]$ into a single UR would allow enough compression to justify the cost of optionality, while the slight compression gained by eliminating a single vowel i from $/paril/$ would not.

We will now show that our learner converges on the correct G_{target} , to which the MDL evaluation metric assigns the best score. Moreover, it will do so without being told which forms (if any) should be collapsed. Since our intention in this subsection is to present a proof-of-concept learning of restricted optionality, we will deviate from our earlier setting and provide the learner with the final constraint set in advance.¹⁷

(15) a. Initial grammar:

$$G_{initial} = \begin{cases} \text{LEX:} & \text{tabil, tab, paril, tapil, tap, radil, labil, lab} \\ \text{CON:} & \text{FAITH} \gg \text{DEP}([-high]) \gg \text{MAX}([-liquid]) \gg * [+cons][+cons] \end{cases}$$

$$\text{Description length: } |G_{initial}| + |D|G_{initial}| = 570 + 600 = 1,170$$

b. Final grammar:

$$G_{final} = \begin{cases} \text{LEX:} & \text{tabl, paril, tapl, radil, labl} \\ \text{CON:} & * [+cons][+cons] \gg \text{FAITH} \gg \text{DEP}([-high]) \gg \text{MAX}([-liquid]) \end{cases}$$

$$\text{Description length: } |G_{final}| + |D|G_{final}| = 396 + 750 = 1,146$$

Significantly, the overgenerating G_{simple} presented above would lead to a longer description length compared to the correct hypothesis: as shown in (16), the grammar itself would be more economical, but the overall description length would be higher.

(16) Overgenerating grammar:

$$G_{simple} = \begin{cases} \text{LEX:} & \text{tabl, parl, tapl, radl, labl} \\ \text{CON:} & * [+cons][+cons] \gg \text{FAITH} \gg \text{DEP}([-high]) \gg \text{MAX}([-liquid]) \end{cases}$$

$$\text{Description length: } |G_{simple}| + |D|G_{simple}| = 368 + 800 = 1,168$$

¹⁶In Dell's original paper, only hypotheses corresponding to G_{simple} and G_{target} are considered. Dell proposes a learning strategy that always favors grammars that are more restrictive, and this strategy works well for cases in which these are the only choices. As we have seen, however, such a strategy will not work in a more general setting: it will reject G_{target} in favor of $G_{identity}$ that does not generalize at all.

¹⁷To keep with our previous assumptions, the initial ranking of the constraints will be a faithful one. Note that l deletion and i epenthesis are equally harmonic. $|D|G|$ is multiplied by 25.

3.4 Alternations

In previous examples we considered phonological grammars that map URs into SRs regardless of the contexts they appear in. Our next step will be to show that our learning criterion extends naturally to learning morphophonological alternations. We will examine the behavior of our learner on a dataset created by concatenating a suffix to a base set of forms. A phonological process would change some of those forms at the boundary, resulting in a subset that is realized differently in two different contexts: members of this subset would surface faithfully when occurring independently, but would be phonologically altered in the environment of the suffix. To illustrate the procedure, consider the Hebrew verbs *katav* (write) and *daag* (worry) along with the 2nd person feminine suffix $-t$. Assuming that Hebrew speakers’ obstruents assimilate in voicing to a following obstruent, our dataset would include *katav*, *kataft*, *daag*, and *daakt*. Our learner will know neither about the morphological constituency of these forms nor that pairs of them are derivationally related. Instead, we will allow the learner to perform segmentation and represent suffixes as part of the lexicon. In addition, following the lead of Goldsmith (2001), lexical items will be allowed to be stored with pointers to affixes that they attach to. If our view of learning as compression is correct, morphophonological alternations should fall off as by-products of two distinct mechanisms: phonological induction, which we have seen in previous sections, and segmentation, which we will now introduce. Thus, if the learner is provided with enough examples, a grammar like the following, presented schematically, should lead to a shorter description length compared to a naive grammar that memorizes the data and captures no generalizations:

(17)

$$G = \begin{cases} \text{LEX:} & \textit{katav}_{\{-t\}}, \textit{daag}_{\{-t\}}; \text{ Suffixes: } \{t\} \\ \text{CON:} & \text{assimilation-triggering constraints} \end{cases}$$

In other words, compressing the lexicon by collapsing multiple SRs into a single UR would justify, in terms of total description length, the addition of assimilation-triggering constraints to CON along with their appropriate ranking. Let us see how this prediction is borne out, using a small dataset of eight words, generated according to the procedure described above.¹⁸

- (18) 1) *katav* 3) *daag* 5) *rakad* 7) *takaf*
 2) *kataft* 4) *daakt* 6) *rakadet* 8) *takaft*

Here, four basic words have been concatenated with a suffix $-t$, triggering two phonological processes. In 1-2 and 3-4, suffixation results with regressive obstruent devoicing. In 5-6, two adjacent coronals are separated by *e* epenthesis, thus blocking voicing assimilation. In 7-8, none of the two environments is met and the basic form remains unchanged. The learner’s task in this case, then, is threefold: to discover the $-t$ suffix by performing segmentation; to learn a constraint ranking that enforces

¹⁸As in section 3.3, we provide the learner with the final constraint set in advance for the present simulation. We also do not incorporate the costs of suffixes and pointers to them. See Goldsmith (2001) for much relevant discussion of how such costs can be taken into account.

regressive devoicing and epenthesis between coronal consonants; and to collapse pairs of SRs into a single UR, without knowing in advance which forms should be collapsed. In the results presented below, all three goals have been reached.

(19) a. Initial grammar:

$$G_{initial} = \left\{ \begin{array}{l} \text{LEX: } \textit{katav}, \textit{daag}, \textit{rakad}, \textit{takaf}, \textit{kataft}, \textit{daakt}, \textit{rakadet}, \\ \textit{takافت}; \text{ Suffixes: } \{\} \\ \text{CON: } \text{FAITH} \gg \text{MAX}([+cons]) \gg \text{DEP}([-ATR]) \gg \text{IDENT}([-voice]) \gg \\ \gg \text{IDENT}([+cons]) \gg \text{IDENT}([+labial]) \gg \text{IDENT}([-labial]) \gg \\ \gg \text{IDENT}([-high]) \gg \text{IDENT}([+high]) \gg *[-coronal][+ATR] \gg \\ \gg * [+coronal][+coronal] \gg * \left[\begin{array}{c} +cons \\ +voice \end{array} \right] [-voice] \end{array} \right.$$

Description length: $|G_{initial}| + |D|G_{initial}| = 864 + 2,400 = 3,264$

b. Final grammar:

$$G_{final} = \left\{ \begin{array}{l} \text{LEX: } \textit{katav}_{[-t]}, \textit{daag}_{[-t]}, \textit{rakad}_{[-t]}, \textit{takaf}_{[-t]}; \text{ Suffixes: } \{t\} \\ \text{CON: } * \left[\begin{array}{c} +cons \\ +voice \end{array} \right] [-voice] \gg *[-coronal][+ATR] \gg \\ \gg * [+coronal][+coronal] \gg \text{IDENT}([-high]) \gg \text{IDENT}([-voice]) \gg \\ \text{DEP}([-ATR]) \gg \text{FAITH} \gg \text{IDENT}([+labial]) \gg \text{MAX}([+cons]) \gg \\ \text{IDENT}([-labial]) \gg \text{IDENT}([+cons]) \gg \text{IDENT}([+high]) \end{array} \right.$$

Description length: $|G_{final}| + |D|G_{final}| = 520 + 1,600 = 2,120$

4 Previous proposals

In the previous sections we developed a proposal for learning in OT. We started by constructing an evaluation metric as part of a discovery procedure for a working phonologist. We saw how compression provided a unifying framework for balancing economy and restrictiveness. We then motivated the use of the very same evaluation metric as a model of the child learner despite the differences – mostly in hypothesis space and in the availability of negative evidence – between the phonologist and the child. We explained why we thought that the evaluation metric is (almost) the bare minimum learner: it is what the child has by virtue of having UG, along with the ability to entertain at least one additional hypothesis at any given time and the ability of traversing the hypothesis space in an appropriate manner. If this reasoning is correct, any deviation would need empirical justification. We proceeded to present a proof-of-concept demonstration of our learner across different datasets, showing its successful learning with *ab-nese*, a toy language with English-like aspiration, a language with restricted optionality, and a morphophonological pattern.

In the present section we will use the perspective provided by the evaluation metric to take a critical look at previous learning algorithms that have been proposed in the

literature on OT. We start, in section 4.1, by briefly reviewing the main efforts in the literature, efforts that, as we will explain, have a somewhat different focus from our own. The next two sections concern proposals that are much closer to our own: Maximum-Likelihood Learning of Lexicons and Grammars (MLG; Jarosz, 2006b,a, 2010), discussed in section 4.2; and Lexical-Entropy Learner (LEL; Riggle, 2006), discussed in section 4.3. We will see that each proposal targets one of the two criteria of economy and restrictiveness but not both, leading to challenges of the kind discussed above for the scientist. Our own proposal, presented earlier, can thus be seen as subsuming both, balancing in a principled way between the two biases.

4.1 Constraint re-ranking approaches

As mentioned in the introduction, OT has spawned a lively discussion of learning. Obviously, we will not be able to do justice to all the relevant literature within the scope of this paper. For the most part, however, this literature has concerned itself with questions that are quite different from those motivating the present proposal. Specifically, some of the most influential proposals, such as Recursive Constraint Demotion (RCD; Tesar and Smolensky, 1998, 2000), Biased Constraint Demotion (BCD; Prince and Tesar, 2004), the Gradual Learning Algorithm (GLA; Boersma and Hayes, 2001), and the Maximum Entropy model of Goldwater and Johnson (2003), assume that the learner has access to pairs of URs and surface forms (as well as a finite inventory of universal constraints). Clearly, these works do not suppose that the child is given these pairs explicitly by the environment. Rather, we are to think of such proposals as part of a bigger system that includes also a learner for the pairings of URs and surface forms. Since it is complete learners that we are interested in, we hope that for the time being it is reasonable to set aside proposals of this kind that rely on an unspecified learner to obtain pairings of URs and surface forms.

Let us now turn to a family of proposals, which we will refer to as paradigm-based lexicon learners, in which constraint re-ranking is combined with some lexical learning. These proposals, which include Tesar (2006, 2008), Apoussidou (2007), Merchant (2008), and Akers (2012), have the following in common: they all use paradigms to extract information about alternations, which in turn support the learning of properties of URs.

Alternations are a central source of information, and we agree with the paradigm-based approach that this source should not be overlooked. For example, it is hard to think of a different basis for learning that the UR for ‘wheel’ in German is *rad* while that of ‘council, advice’ is *rat* : the surface form in both cases is *[rat]*; but the plural form of ‘wheel’ is *rade*, while that of ‘council’ is *rete*. However, while alternations are undoubtedly important in discovering URs, they are a special case of a more general phenomenon and would ideally fall out of whatever mechanism takes care of handling aspiration in English, a phenomenon for which alternations offer no help. The MDL learner that we presented above treats alternation-based learning as exactly this kind of special case, as we have tried to show in section 3.4. Paradigm-based learners, on the other hand, treat alternations as a world unto itself. Not surprisingly, then, the paradigm-based learners in the literature offer no obvious generalization for properties of URs that do not come from alternations, leaving us no closer to discovering the

facts about aspiration in English, for example, than we were before. Until a paradigm-based learner is proposed that generalizes beyond alternations, we conclude that, like the constraint re-ranking approaches mentioned above, such learners can be set aside within the context of the present discussion.

4.2 Maximum-Likelihood Learning of Lexicons and Grammars

Jarosz (2006b,a) proposes an algorithm for learning lexicons and constraint rankings based on the principle of Maximum Likelihood (ML). Working within a probabilistic version of OT, Jarosz assumes that a hypothesis is a distribution over constraint rankings coupled with a distribution over URs for each morpheme.¹⁹ The learner is given the set of constraints in advance (either as part of the innate component or perhaps through a separate module for learning constraints), along with candidate URs for each morpheme. The learner then attempts to find the hypothesis that maximizes the likelihood of the data. The search starts with an uncommitted lexicon, in which all candidates for any given morpheme are equally likely, and the search for the best hypothesis is performed by the Expectation Maximization algorithm (EM; Dempster et al., 1977).

Let us demonstrate with a simple variation on *ab*-nese in which we have the same input sequence as in (1) above but are restricted to working with the constraints **ab*, **p*, and IDENT, all three of which are given to us in advance; we will also assume the knowledge that *b* has *p* as a featural variant and that *a* has *e*. The learning process will start with the hypothesis that for any given morpheme, all possible URs are equally likely. That is, the initial hypothesis provides the following distribution over the lexicon (along with a distribution over the possible rankings of the constraints):

- (20) a. M_1 =(*ab*) URs: *ab* (.25); *ap* (.25); *eb* (.25); *ep* (.25)
 b. M_2 =(*bab*) URs: *bab* (.125); *bap* (.125); *beb* (.125); *pab* (.125); *bep* (.125); *pap* (.125); *peb* (.125); *pep* (.125)
 c. M_3 =(*aaab*) URs: *aaab* (.0625); *aaap* (.0625); *aaeb* (.0625); *aeab* (.0625); *eaab* (.0625); *aeap* (.0625); ...
 d. M_4 =(*aabab*) URs: ...
 e. M_5 =(*baab*) URs: ...
 f. M_6 =(*aababaaaabab*) URs: ...
 g. M_7 =(*babababaa*) URs: ...
 h. M_8 =(*babababababababab*) URs: ...

On Jarosz's assumptions, the correct morpheme for each surface form has been identified in advance and is available to the learner. Using this knowledge, each hypothesis defines a probability distribution over surface forms that can be computed by enumerating the possible URs and the different constraint rankings. Take the surface form *ab*, for example: suppose we encounter it in a certain position in the data, and

¹⁹This probabilistic version of OT is distinct from Stochastic OT (Boersma, 1998; Boersma and Hayes, 2001).

suppose further that this position has been correctly identified as expressing the morpheme M_1 . Our goal is to compute its likelihood, and we do this by enumerating the different URs that M_1 is associated with – in this case, ab , ap , eb , and ep – and by computing the conditional probability of the surface form ab given each of the URs; the final answer is the weighted sum of these conditional probabilities, each weighted by the probability of the relevant UR:

(21) The likelihood of the surface form ab given that the morpheme is M_1

$$P(\text{surface} = ab | M_1) = \sum_{u \in \{ab, ap, eb, ep\}} P(\text{surface} = ab | u) P(u)$$

The probabilities of the different URs are part of each hypothesis. For example, in the initial hypothesis (20), the distribution is uniform, with each UR for M_1 having a probability of 0.25. What remains is the computation of $P(\text{surface} = ab | u)$ for any particular UR. This is done by looking at the different constraint rankings and their probabilities (again, part of every hypothesis). Let us look at how this is done for the UR ab :

(22)

Hypothesis H			Probability under input ab		
Ranking r_i		$P(r_i)$	Optimal O_k		
r_1	$*ab \gg *p \gg \text{IDENT}$	0.2	eb		
r_2	$*ab \gg \text{IDENT} \gg *p$	0.15	eb		
r_3	$\text{IDENT} \gg *ab \gg *p$	0.05	ab		
r_4	$*p \gg *ab \gg \text{IDENT}$	0.1	eb		
r_5	$*p \gg \text{IDENT} \gg *ab$	0.0	ab		
r_6	$\text{IDENT} \gg *p \gg *ab$	0.5	ab		

The probability of the surface form ab given the UR ab is obtained by summing over the rows in which the surface form ab is the winner. In the present case, these are the third, fifth, and sixth rows: $P(r_3) + P(r_5) + P(r_6) = 0.05 + 0.0 + 0.5 = 0.55$. By repeating the computation with the other possible URs for M_1 we obtain the required values to compute the likelihood of the surface form ab given M_1 according to (21).

ML addresses the restrictiveness requirement directly: any overgeneration will lead to spending probability mass on forms that do not occur.²⁰ An ML grammar is thus a fully restrictive one. Meanwhile, starting from an uncommitted lexicon as in (20) encourages the learner to consider hypotheses that rely on the constraints – rather than on accidents of the lexicon – to encode patterns in the input. Such hypotheses accord well with the OT principle of ROTB, mentioned earlier, which states that the set of inputs is universal. From an information-theoretic perspective, an uncommitted lexicon is one with high entropy. As we will see in the next section, lexicon entropy (though in different form from Jarosz’s) can sometimes stand proxy for economy, both criteria sometimes favoring a smaller lexicon from which significant patterns have been extracted over a more complex one in which those patterns remain.

As we saw earlier, restrictiveness must be simultaneously balanced against economy in order to provide an adequate evaluation of hypotheses. Above we discussed the

²⁰This closely mirrors the minimization of $|D|G|$ alone under a description-length approach.

word. As we discussed, a restrictiveness-only phonologist will fail to learn this generalization, memorizing instead the finite subset of the infinite allowable z -forms seen so far. The MLG, similarly aiming for restrictiveness only, will run into the very same problem.

What about the uncommitted initial state as a cure for memorization? In Jarosz's examples, the learner does not end up memorizing the input, and we mentioned that the uncommitted initial state of the MLG is designed to encourage the learner to be reasonable. Unfortunately, such encouragement is generally short lived. It affects the beginning of the search, but if the search is capable enough, the ML criterion will necessarily lead the learner to a maximally memorized hypothesis. That Jarosz's examples do not exhibit such memorization we must attribute to peculiarities of the search procedure. The EM algorithm is known to get caught in local optima, which could account for these results. Moreover, it is possible that the search has stopped before convergence. Since modeling the search goes beyond the goals of current research, we conclude that the entropic initial state is not capable of rescuing ML.

4.3 Lexical-Entropy Learner

We just saw that an uncommitted – or entropic – initial state does little to help the learner. Assuming that an entropic lexicon is indeed a relevant property of good hypotheses, the remedy seems clear: turn the requirement into an active force by incorporating it into the learning criterion. This is exactly what Riggle (2006) proposes. On Riggle's account, different grammar and lexicon hypotheses are evaluated according to a measure of lexicon entropy. The measure, which is somewhat different from Jarosz's and which we will discuss shortly, is based on the following principle:

- (25) Lexicon entropy principle. Assume a universal constraint set CON. Whenever faced with a decision whether to encode a phonological pattern as a consequence of constraint interaction or as an accident of the lexicon the former option must be taken. (Modified from Riggle, 2006, p. 347.)

Riggle proposes to evaluate the lexicon's entropy by computing mutual information for bigrams, according to the following function:

$$(26) \quad H(Y|X) = - \sum_{x \in \Sigma} \sum_{y \in \Sigma} P(x, y) \log P(y|x)$$

As an example of how this should work, consider again the *ab*-nese data from section 2.1, and the three constraints **bb*, MAX, and DEP. We will show why Riggle's metric rejects the identity hypothesis in favor of the correct lexicon and ranking combination. The data are repeated here, along with the two competing hypotheses:

- (27) Hypothesis 1 (identity)

Lexicon:

- | | | | |
|----------|------------|--------------|----------------------|
| 1) /ab/ | 3) /aaab/ | 5) /abaab/ | 7) /babababaa/ |
| 2) /bab/ | 4) /aabab/ | 6) /babaaaa/ | 8) /bababababababab/ |

Corresponding ranking: any

Entropy: 0.63

(28) Hypothesis 2 (correct)

Lexicon:

- 1) /ab/ 3) /aaab/ 5) /abaab/ 7) /bbbbaa/
2) /bb/ 4) /aabb/ 6) /bbaaaa/ 8) /bbbbbb/

Corresponding ranking: **bb*, MAX \gg DEP

Entropy: 1.55

The lexicon of hypothesis 1 is identical to the surface data. Under any ranking of the three constraints, all underlying representations would surface unchanged. The generalization that a sequence *bb* is prohibited in *ab*-nese is captured only as an accident of the lexicon. As a consequence, the lexicon contains predictable information that can be identified by computing probabilities of adjacent segments: after seeing a *b*, there is a probability of 1.0 that a following segment be *a*. Formally, $-P(b, b) \log P(b|b)$ and $-P(b, a) \log P(a|b)$ will both be null (assuming here for simplicity's sake that $0 \log 0 = \lim_{x \rightarrow 0} x \log x = 0$), not adding to the entropy of the lexicon, which results in 0.63.

On the other hand, the second hypothesis has the predictable information about the absence of *bb* sequences removed from the lexicon, resulting in a lexicon that contains no regularities: seeing a consonant or a vowel, it is hard to predict what the next segment would be. Here, all summands contribute to the measure of entropy, which sums to 1.55 — a higher entropy than that of the identity hypothesis. Importantly, given the lexicon of the second hypothesis, a sequence *bb* must be resolved by vowel epenthesis, entailing the more restrictive ranking **bb* \gg DEP.

We can see that Riggle uses entropy as a proxy for economy.²³ In his proposal, entropy is the only factor in the learning criterion. In particular, there is no pressure for restrictiveness. This choice leads to the subset problem, the problem discussed earlier for the scientist using the original SPE evaluation metric and the mirror image of the problem for Jarosz's proposal. To see this, let us consider first a version of Riggle's proposal for *ab*-nese in which the constraints are not given in advance and must be learned. In the absence of a pressure for restrictiveness, an entropic but overgenerating grammar like the following will fare much better than the correct grammar:

(29) Hypothesis 3 (no constraints; entropic; overgenerates)

Lexicon: /aabba/

Corresponding ranking: (NONE: no constraints to rank)

Entropy: 0.63

In Lexicon 3, the bigram distribution is uniform: $P(x|y)$ is the same ($= 0.5$) for any *x* and *y*. It is thus maximally entropic. At the same time, it massively overgenerates: in the absence of any constraints, the single UR /aabba/ can be mapped to any of the attested forms but also to any other form, all without incurring a single violation. In this case, then, entropy alone exposes the learner to the subset problem, just as economy alone exposed the scientist to this problem in our discussion earlier.

²³Note that Riggle's version applies to lean lexicons rather than the rich lexicons of Jarosz's system. In this respect, Jarosz's proposal demonstrates stricter adherence to the letter of ROTB than Riggle does.

In Riggle’s actual proposal, the constraints are given to the learner in advance. With a judicious choice of constraints, the subset problem is ameliorated, but we will show that it does not disappear completely. Let us continue with our *ab*-nese example, and let us assume that the learner is given the set of constraints that our phonologist from our earlier discussion arrived at. In this case, Lexicon 3 is no longer appropriate (since it does not generate the data), but the following overgenerating hypothesis is equally entropic as the intuitively correct Hypothesis 2:

(30) Hypothesis 4 (overgenerates)

Lexicon:

- | | | | |
|---------|-----------|-------------|-------------|
| 1) /ab/ | 3) /aaab/ | 5) /abaab/ | 7) /bbbbaa/ |
| 2) /bb/ | 4) /aabb/ | 6) /bbaaaa/ | 8) /bbbbbb/ |

Corresponding ranking: $*bb \gg \text{MAX}, \text{DEP}$

Entropy: 1.55

Hypothesis 4 keeps the ranking $*bb \gg \text{DEP}$, but it has MAX ranked together with DEP rather than above it. As a result, all the correct surface forms are still generated from the intuitively correct lexicon, but along with them we will also find unattested forms such as *b* (from the UR *bb*), generated through *b*-deletion. Since lexicon entropy does not take restrictiveness into account, such overgeneration will not lead to Hypothesis 4 being dispreferred.²⁴

In order to assess the suitability of entropy as a pressure on hypotheses, then, we must combine it with a pressure for restrictiveness. The natural way to accomplish this is by combining it with Jarosz’s ML criterion. There are many different ways to combine two criteria into one, and many of these (such as maximizing the sum – or the product – of the likelihood of the data and the entropy of the lexicon) will address the problem of overgeneration without degenerating into memorizing the input.

Unfortunately, no combination of this kind can work. To see why, consider again the two lexicons for *ab*-nese that seemed to justify the entropy criterion. Lexicon 1 was more complex and less entropic than Lexicon 2, which seemed encouraging. But consider now Lexicon 5, a lexicon based on *c*-deletion rather than *a*-insertion:

(31) Hypothesis 5 (entropic and restrictive; presumably bad)

Lexicon:

- | | | | |
|------------|--------------|--------------|----------------------|
| 1) /cacbc/ | 3) /caaaccb/ | 5) /abaab/ | 7) /babababaa/ |
| 2) /bab/ | 4) /aabab/ | 6) /babaaaa/ | 8) /bababababababab/ |

Corresponding ranking: $*c, *bb, \text{DEP} \gg \text{MAX}$

Entropy: 1.578

²⁴Allowing MAX and DEP to be ranked together is in line with certain variants of OT – see, in particular, Anttila (2007), who argues for the use of such rankings to account for optionality; a similar state of affairs is also possible within Stochastic OT (Boersma, 1998; Boersma and Hayes, 2001) – but we have chosen it here simply to make the presentation of the current point easier. We could have made the same point while adhering to strict linear orderings of the constraints, for example by considering a variant of *ab*-nese in which the following hold: two occurrences of *b* in a row are okay; three are not; an underlying *bbb* sequence can be repaired by a single insertion of *a* after the first occurrence of *b* but not after the second. A correct grammar would enforce the positional requirement on the insertion of *a*. For Riggle, however, the ranking $*bbb \gg \text{MAX} \gg \text{DEP}$ will do just as well, despite the fact that it overgenerates by allowing an underlying *bbb* sequence to surface both (correctly) as *babb* and (incorrectly) as *bbab*.

Lexicon 5 is more complex still than Lexicon 1, but it is more entropic than either Lexicon 1 or Lexicon 2. In fact, infinitely many such lexicons are easily constructed. Note that all the hypotheses in this case are fully restrictive, so ML will not help choose between them. The decision is down to entropy, and entropy leads us astray: it only cares about making the grammar less regular, but this can be accomplished not just by removing orderly material, which is what we would like, but also by adding disorderly material, which we would not. We conclude that economy must be represented directly, as it is under MDL, rather than by proxy.

5 Discussion

The evaluation metric of SPE expressed a hope: let the theoretical linguist focus on building the right theory of UG; given that theory, the evaluation metric will take the child from the initial state to adult knowledge using very general considerations and the data at hand. Learning, in a sense, will take care of itself.

In the years that followed, that hope came to seem increasingly naive. The evaluation metric foundered on the subset problem: any metric focusing exclusively on economy would. The challenge of negotiating an infinite hypothesis space did little to help. By the time OT arrived on the scene, the evaluation metric was no longer actively pursued.

That original hope may have been abandoned too readily, as we have tried to show in this paper. The particular choice of the SPE metric was problematic, but the MDL alternative from Solomonoff's work addresses the challenges to the SPE metric in a fully general way. At first glance, the compression criterion at the heart of Solomonoff's metric can seem foreign from the perspective of linguistics. We tried to show, however, that this criterion is in fact familiar from the everyday work of the phonologist. We then presented the case for using this criterion as the null hypothesis about the child's learning criterion: given any theory of UG, the ability to store grammars in memory and use them to parse the data already provides the basis for using the MDL metric; anyone who wishes to argue that the child is prevented from using this null criterion and instead uses some different learning method would need to provide supporting evidence. We proceeded to present several simulation results showing how the MDL metric can be used by a learner trying to make sense of raw data. While clearly preliminary, these proof-of-concept results – all of them new – show how lexicons can be induced, with and without supporting data from alternations, and how the same metric extends to learning the constraints themselves.

References

- Akers, Crystal Gayle. 2012. Commitment-based learning of hidden linguistic structures. Doctoral Dissertation, Rutgers University-Graduate School-New Brunswick.
- Albright, Adam, and Bruce Hayes. 2011. Learning and learnability in phonology. In *The handbook of phonological theory*, ed. John Goldsmith, Jason Riggle, and Alan Yu, 661–690. Wiley Online Library, 2nd edition.

- Anttila, Arto. 2007. Variation and optionality. *The Cambridge handbook of phonology* 519–536.
- Apoussidou, Diana. 2007. *The learnability of metrical phonology*. LOT.
- Baker, C. L. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10:533–581.
- 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.
- Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Holland Academic Graphics/IFOTT.
- Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.
- Braine, M. D. S. 1971. On two types of models of the internalization of grammars. In *The ontogenesis of grammar*, ed. D. J. Slobin, 153–186. Academic Press.
- Chaitin, Gregory J. 1966. On the length of programs for computing finite binary sequences. *Journal of the ACM* 13:547–569.
- Chater, Nick, and Paul Vitányi. 2007. ‘Ideal learning’ of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology* 51:135–163.
- Chomsky, Noam. 1951. Morphophonemics of Modern Hebrew. Master’s thesis, University of Pennsylvania.
- Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.
- 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.
- Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12:31–37.
- Dempster, Arthur Pentland, 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)* 39:1–38.
- Dowman, Mike. 2007. Minimum description length as a solution to the problem of generalization in syntactic theory. Ms., University of Tokyo, Under review.
- Endress, Ansgar, Ghislaine Dehaene-Lambertz, and Jacques Mehler. 2007. Perceptual constraints and the learnability of simple grammars. *Cognition* 105:577–614.
- Endress, Ansgar, Marina Nespors, and Jacques Mehler. 2009. Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences* 13:348–353.
- Endress, Ansgar D, and Jacques Mehler. 2010. Perceptual constraints in phonotactic learning. *Journal of experimental psychology. Human perception and performance* 36:235–250.
- Gibson, Edward, and Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25:407–454.
- Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:153–198.
- Goldsmith, John. 2010. Towards a new empiricism for linguistics. To appear as chapter

- 3 in *Empiricist Approaches to Language Learning*, co-authored with Alex Clark, Nick Chater, and Amy Perfors.
- Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–120.
- 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.
- Hale, Mark, and Charles Reiss. 1998. Formal and empirical arguments concerning phonological acquisition. *Linguistic Inquiry* 29:656–683.
- Halle, Morris. 1962. Phonology in generative grammar. *Word* 18:54–72.
- Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.
- Heinz, Jeffrey. 2007. The inductive learning of phonotactic patterns. Doctoral Dissertation, University of California, Los Angeles.
- Heinz, Jeffrey, Gregory Kobele, and Jason Riggle. 2009. Evaluating the complexity of Optimality Theory. *Linguistic Inquiry* 40:277–288.
- Hsu, Anne S., and Nick Chater. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science* 34:972–1016.
- Jarosz, Gaja. 2006a. Rich lexicons and restrictive grammars – Maximum Likelihood learning in Optimality Theory. Doctoral Dissertation, Johns Hopkins University, Baltimore, Maryland.
- Jarosz, Gaja. 2006b. 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. 2010. Naive parameter learning for Optimality Theory – the hidden structure problem. Ms.
- Kiparsky, Paul. 2007. Panini’s razor. Slides from a talk given at the First International Sanskrit Computational Linguistics Symposium, October 2007.
- Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220:671–680.
- Kolmogorov, Andrei Nikolaevic. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission (Problemy Peredachi Informatsii)* 1:1–7. Republished as ?).
- Li, Ming, and Paul Vitányi. 2008. *An introduction to kolmogorov complexity and its applications*. Berlin: Springer Verlag, 3rd edition.
- Manzini, M. Rita, and Kenneth Wexler. 1987. Parameters, binding theory, and learnability. *Linguistic Inquiry* 18:413–444.
- de Marcken, Carl. 1996. Unsupervised language acquisition. Doctoral Dissertation, MIT, Cambridge, Mass.
- Merchant, Nazarré Nathaniel. 2008. Discovering underlying forms: Contrast pairs and ranking. Doctoral Dissertation, Rutgers, The State University of New Jersey.
- Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25:83–127.

- Niyogi, Partha, and Robert C. Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61:161–193.
- Peña, Marcela, Luca Bonatti, Marina Nespor, and Jacques Mehler. 2002. Signal-driven computations in speech processing. *Science* 298:604–607.
- 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.
- Riggle, Jason. 2006. Using entropy to learn OT grammars from surface forms alone. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 346–353.
- 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.
- Smith, Kirk H. 1966. Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology* 72:580–588.
- 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. 1960. A preliminary report on a general theory of inductive inference. Technical Report ZTB-138, Zator Co., Cambridge, MA.
- Solomonoff, Ray J. 1964a. A formal theory of inductive inference, part I. *Information and Control* 7:1–22.
- Solomonoff, Ray J. 1964b. A formal theory of inductive inference, part II. *Information and Control* 7:224–254.
- Stolcke, Andreas. 1994. Bayesian learning of probabilistic language models. Doctoral Dissertation, University of California at Berkeley, Berkeley, California.
- Tesar, Bruce. 2006. Faithful contrastive features in learning. *Cognitive Science* 30:863–903.
- Tesar, Bruce. 2008. Learning phonological grammars for output-driven maps. To appear in NELS 39.
- 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, MA: MIT Press.
- Wallace, C.S., and D.M. Boulton. 1968. An information measure for classification. *Computer Journal* 11:185–194.