Learning alternations from surface forms with sublexical phonology[*]

May 19, 2015
comments welcome

Blake Allen
University of British Columbia
b.allen@alumni.ubc.ca

Michael Becker
Stony Brook University
michael.becker@phonologist.org

Abstract

We present a constraint-based multiple-grammar approach to learning and productively applying morphophonological alternations, with a focus on non-local and non-concatenative phenomena, based on the *sublexical* approach (Becker & Gouskova to appear). Success of the model is demonstrated on the English plural and English past tense data sets, among a variety of others. We compare our model to the rule-based Minimal Generalization Learner, showing our model's ability to handle a wider range of morphological patterns. Our model contrasts with constraint-based proposals that use abstract underlying representations to generate surface forms in a single grammar; we show that shifting the locus of complexity from the representations to the grammar allows learning from realistically large and noisy data.

Keywords: morphophonology, exceptions, variation, grammatical vs. representational, underlying representations, sublexical phonology.

[*] For their thoughtful comments and suggestions, we thank Adam Albright, Ellen Broselow, Robert Daland, Naomi Feldman, Bruce Hayes, Peter Jurgec, Douglas Pulleyblank, Colin Wilson, Kie Zuraw, and audiences at Phonology 2014, the 89th LSA, NYU and UMD. Special thanks to Maria Gouskova, whose thinking permeates much of this work.

We thank Kie Zuraw for sharing Tagalog data, and Bruce Hayes and Adam Albright for sharing English and Spanish data.

We also thank Peter Kleiweg for sharing the code for the minimal edit distance algorithm, and Ilia Kurenkov and Omer Preminger for help with the Javascript implementation. Remaining errors are the other author's.

1 Introduction

Speakers are able to apply the morphophonological alternations of their language productively to novel words. For example, English speakers given the nonce singular noun [blaɪf] *'blife'* are able to generate corresponding plural forms such as [blaɪfs] and [blaɪvz], assess the relative wellformedness of each plural candidate, and also identify the much lower acceptability of [blaɪfɪz]. This knowledge is largely morpheme-specific, e.g. the plural of [boɹ] 'boar' is necessarily [boɹz], despite the language-wide phonotactic well-formedness of [boɹs] and [boɹɪz] (Berko 1958).

We present a computationally implemented model of this grammatical knowledge with an associated learning algorithm, which together mimic the creation and productive application of morphophonological generalizations. The learning procedure operates on a list of two-member paradigms (such as English singular-plural pairs) and creates a small set of operations and grammars which are then applied to test words, generating output forms and predicting a probability distribution over them. These morphological operations (e.g. "add [z] at the right edge") generate candidate surface forms for the plural directly from the surface form of the singular. There is no search for underlying representations, and no further manipulation of the plural once it is formed. Each operation is paired with an operation-specific grammar that governs its distribution, e.g. limiting the addition of [ɪz] to strident-final nouns. Following Becker & Gouskova (to appear), we use the term *sublexicon* for each operation-grammar pair.

We focus our attention on realistically large and noisy data sets, in some cases including thousands of word pairs, following the example of the Minimal Generalization Learner (MGL, Albright & Hayes 2002, 2003, 2006). Unlike the MGL's reliance on a proliferation of rules that combine structural changes and environments, we use a relatively small number of constraint-based grammars that separate the observed morphological changes (e.g. "add [ɪz]") from the phonological conditions on them (e.g. having a word-final strident). This separation allows the model to learn the location of multiple morphological changes and their potentially non-local predictors, as discussed below.

We contrast our approach with previous constraint-based work on morphological learning built on the premise of a single grammar and a search for underlying representations, e.g. Tesar & Smolensky (1998, 2000); Tesar et al. (2003); Jarosz (2006); Riggle (2006); Rasin & Katzir (2013). While these approaches have

	content	position	distribution
a.	[s]	right edge	after non-strident voiceless segments
b.	[z]	right edge	after non-strident voiced segments
c.	[ɪz]	right edge	after strident segments

Table 1: The English regular plural

undeniable merits, they are limited by their goal of learning a single grammar for a given language while only using morpheme-blind markedness and faithfulness constraints. Theoretical work in constraint-based theory regularly goes beyond a single language-wide morpheme-blind grammar, starting with the very work that introduced Optimality Theory (Prince & Smolensky 1993/2004), which included Align(um), a morpheme-specific constraint for the analysis of Tagalog. The many approaches that use multiple grammars and/or morpheme-specific constraints include cophonologies (Inkelas et al. 1996; Inkelas & Zoll 2007; Anttila 2002, a.o), constraint indexation (Pater 2000, 2006, 2008; Fukazawa 1999; Itô & Mester 1995, 1999; Kawahara et al. 2002; Flack 2007; Gouskova 2007; Becker 2009; Becker et al. 2011, a.o.), UseListed (Zuraw 2000; Hayes & Londe 2006; Becker et al. 2012), and a variety of others. These works all implicitly or explicitly reject the prospect of a single language-wide morpheme-blind grammar.

The paper is organized as follows. We first present the learner using the example of the English plural suffix in §2. We then show how the learner learns a typologically diverse range of morphology in §3. The model's ability to handle the complexity of the English past tense is demonstrated in §4, and §5 concludes.

2 Learning local alternations: the English plural

The regular English plural is a familiar example of a relatively simple case of affixation. The vast majority of English nouns form their plural as seen in Table 1, e.g. [dʌk] 'duck' forms its plural via the addition of [s] at the right edge due to its final segment being a voiceless non-strident. Similarly, [z] is added after voiced non-stridents, and [ɪz] after stridents. In this section, we demonstrate the Sublexical Learner's method for discovering the content and location of the plural marker's various realizations as well as the phonological predictors of their distributions.

Discovering the content and position of the plural exponents is discussed in

b	ɹ	ʌ	ʃ	∅	∅
b	ɹ	ʌ	ʃ	ɪ	z

Figure 1: Alignment of English [bɹʌʃ ∼ bɹʌʃɪz] ‘brush(es)’

§2.1, their distribution is discussed in §2.2, and application to nonce forms is discussed in §2.3. We turn to a more realistic view of the English plural that includes stem alternations in §2.4. We summarize and compare to representational approaches in §2.5.

A browser-based implementation of our learner and all of the simulations discussed in this paper are available at `http://sublexical.phonologist.org/`, which also features a link to the learner’s source code.

2.1 Learning exponent content and positions

The input to the learner is a list of pairs of words, supplied in their unanalyzed surface forms. We call each pair of forms a *paradigm*, each of which comprises a *base* form and a *derivative* form. We assume that learning proceeds once the human learner has conceptually connected each pair of words, for example singular nouns and their plurals (cf. MDL-based efforts to establish these connections, e.g. Goldsmith 2001; Rasin & Katzir 2013). The learner also requires a feature matrix that lists all segments and their corresponding phonological feature values. Features are used early in learning to improve paradigm alignments and then again later to assess constraint violations (see §2.2).

For each base–derivative pair, the learner calculates the featural distance of segment pairs and then uses dynamic programming to backtrack over this distance matrix, producing alignments like the one shown in Figure 1 for [bɹʌʃ ∼ bɹʌʃɪz] ‘brush(es)’.[1] Identical or similar segments line up one-to-one, and any unpaired

[1] Using the standard edit distance algorithm, also known as the Levenshtein algorithm or Needleman-Wunsch algorithm, would pair non-identical sounds indiscriminately. Our distance metric is based on Peter Kleiweg’s (Kleiweg 2014), whereby the distance between segments is a function of the number of non-shared phonological feature values between them. Such featurally-improved alignment has been used by Nerbonne et al. 1996; Heeringa 2004; Spruit et al. 2007; Beijering et al. 2008; Nerbonne & Heeringa 2010, a.o. Using featural alignment usually generates desirable alignments of word pairs that exhibit rather stark featural dissimilarities. For example, in the German [bux ∼ byçəɐ̯] ‘book(s)’, it ensures that the [u] aligns with the [y], and the [x] aligns with the [ç]. We will see below, however, that the learner can overcome suboptimal alignments when they do arise. For a different use of minimal edit distance alignments to learn morphophonology based on spelling, see Durrett & DeNero (2013) and references within.

	base	possessed	
a.	bas	bas-**ka**	‘hair’
	sa.na	sa.na-**ka**	‘deer’
	sa.paa	sa.paa-**ka**	‘forehead’
b.	si.wa.nak	si.wa-**ka**-nak	‘root’
	suu.lu	suu-**ka**-lu	‘dog’
	kuh.bil	kuh-**ka**-bil	‘knife’

Table 2: Ulwa: left-oriented [ka] sometimes appears at the right edge.

segments are aligned with null symbols (cf. McCarthy & Wolf 2010).

Taking the first (top) form as the base and the second (bottom) form as the derivative, the learner uses this alignment to extract the morphological operation “add [ɪz]”, thus identifying the content of the operation for this particular paradigm, its *exponent.* In addition to the segmental composition of the affix, the learner must learn its position, i.e. it must learn that when applied to a nonce word, the addition of [ɪz] must occur at that word’s right edge.

We do not conclude that simply because [ɪz] is suffixal in a single paradigm like [bɹʌʃ ∼ bɹʌʃɪz] that it is generally suffixal, even if this is the typologically common case. The need for such caution is exemplified by cases of infixation, e.g. the possessive affix [ka] of Ulwa (Hale & Blanco 1989; Bromberger & Halle 1988; McCarthy & Prince 1993a,b), as seen in Table 2. In this language, [ka] appears at the right edge of the base form for shorter words (Table 2a), but longer words disconfirm this hypothesis about the affix’s position (Table 2b). The correct generalization is that the position of [ka] is computed relative to the left edge, as it appears after the base’s second syllable if it is light but after the base’s first syllable otherwise.

We assume that morphological exponents must be left-oriented or right-oriented, i.e. the position of affixes is computed relative to a word edge, even though an affix does not necessarily appear exactly at the edge (following, e.g. Broselow & McCarthy 1983; McCarthy & Prince 1995, see §3.3, §3.4).

To infer the position of an exponent, the learner takes into account all relevant paradigms in the aggregate. We implement a mechanism that first generates multiple *hypotheses* about the position of the exponent from each paradigm separately and then allows more reliable hypotheses to *consume* similar but less desirable ones, gradually homing in on the most reliable generalization

Hypothesis H_L is consumed by hypothesis H_W iff
H_W and H_L are *idempotent* on H_L's bases
and H_W covers *more lexical items* than H_L.

Table 3: Hypothesis reduction (interim, final version in Table 12)

for each operation. This process eventually arrives at the smallest number of hypotheses which can together generate all observed derivatives (plurals in this case). Formally, a hypothesis is consumed if the set of derivatives it generates is a proper subset of the derivatives generated by some other hypothesis. We provide a definition in Table 3, where *idempotence* refers to the ability of one hypothesis to create the same derivatives as another for a particular base, and coverage is defined as the number of paradigms that a hypothesis can derive correctly.

Hypotheses are either left-oriented or right-oriented, and for the most part they state distance from the edge in terms of segments; see §3.4 for non-segment-counting hypotheses. We use segment-counting as a first step towards a more realistic theory of change loci; see §3.5 for further discussion.

In the English case, the paradigm [bɹʌʃ ∼ bɹʌʃɪz] gives rise to the left-oriented hypothesis "add [ɪz] after the fourth segment" and the right-oriented hypothesis "add [ɪz] at the right edge". This process is repeated for each paradigm in the data set individually, then identical hypotheses are aggregated. Each aggregate hypothesis consists of an operation that states a change and a position (e.g. "add [ɪz] after the final segment") and the list of words that support it (e.g. 'brush', 'glass', and 'quiche'). Once the learner has generated these hypotheses, it moves to the hypothesis reduction step, with the goal of keeping the smallest number of hypotheses able to generate all of the observed derivatives.

To train the learner on a realistic amount of data, we took a sample of 10,000 nouns from the CMU Pronouncing Dictionary (2008, version 0.7a), excluding irregulars and any nouns that end in [f] or [θ] (which we discuss in §2.4). The ten hypotheses from this data set with the greatest coverage are shown in Table 4, where consumption will eliminate all but hypotheses 1, 2, and 7.

Hypothesis consumption has the additional advantage of weeding out undesirable alignments. Such alignments arise when the base and affix share segments, as in [ɹoʊz ∼ ɹoʊzɪz] 'rose(s)'. Here, the minimal edit distance algorithm would ideally produce the alignment in Figure 2a, identifying an affix [ɪz] that can potentially generalize to any strident-final noun; yet it greedily matches the [z] segments at the edges, incorrectly aligning the stem's [z] with the suffix's

	hypothesis	coverage	
1.	**add [z] at right edge**	6689	stʌb, smɑɹti, aɪbɔl, ætəm, …
2.	**add [s] at right edge**	2652	finətaɪp, spɛk, pʊt, əʤʌstmənt,…
3.	add [z] after 5th segment	1319	glɪtɚ, ɑɹtɚi, tɪmbɚ, æfgæn, …
4.	add [z] after 6th segment	1125	smɑɹti, pɑmpɑn, tɛksən, …
5.	add [z] after 4th segment	1099	stʌb, wɚkɚ, lænd, vɛʤi, …
6.	add [z] after 7th segment	855	sɔftbɔl, ɪnkɚʒən, sɚtənti, …
7.	**add [ɪz] at right edge**	762	aɪs, bɹænʧ, lɚʧ, ɔɹɪnʤ, pɹɪvlɪʤ, …
8.	add [z] after 3rd segment	750	fɪn, ɹɪb, læd, bɛl, kɹu, hɚi, …
9.	add [z] after 8th segment	599	kætəpɪlɚ, əflɪkʃən, hʌmɪŋbɚd, …
10.	add [s] after 4th segment	423	spɛk, fɔɹt, bʌŋk, kɹɑp, …

Table 4: Ten largest hypotheses before consumption for the English plural

a.

ɹ	oʊ	z	∅	∅
ɹ	oʊ	z	ɪ	z

b.

ɹ	oʊ	∅	∅	z
ɹ	oʊ	z	ɪ	z

Figure 2: Misalignment due to identity of stem and affix edge segments

[z], as shown in Figure 2b. The leftover segments [zɪ] are identified as an infix, creating a potential generalization that could only be applied to [z]-final nouns rather than all strident-final nouns. This problem is certainly not particular to the English plural, as it would arise in any similar case of shared material between base and affix, such as the Spanish [pais ∼ paises] 'country.SG/PL', or the Hebrew [agam ∼ agamim] 'lake.SG/PL'. Changing direction in an attempt to improve the alignment of suffixes would only serve to worsen the alignment of prefixes in mirror-image cases, e.g. English [ɹoʊl ∼ ɹi-ɹoʊl] '(re)roll', or Slovenian [delati ∼ do-delati] 'work.IMPERF/PERF'.

Given the alignment in Figure 2b, the learner generates two undesirable hypotheses, the left-oriented "add [zɪ] after 2nd segment" and the right-oriented "add [zɪ] before final segment". These hypotheses can only generate grammatical plurals for small regions of English singular forms: three-segment [z]-final singulars and arbitrary-length [z]-final singulars, respectively. Ultimately, both of these undesired hypotheses are consumed by "add [ɪz] after final segment", resulting in a reduction of the hypotheses to the ones in Table 1.

To summarize, then, the procedure we propose for the creation and subsequent consumption of hypotheses allows the learner to identify both the content and positions of affixes directly from the surface forms of the paradigm, with no additional supervision. We have shown this ability in a simple case of suffixation; more complex cases are explored below.

2.2 Learning exponent distributions

We have seen that the learner aggregates words that share a morphological exponent (e.g. "add [s]") and then uses hypothesis reduction to identify their location (e.g. "at the right edge"). The learner needs to learn the phonological generalizations that predict the distributions of the different exponents, which it does by pairing each hypothesis with two constraint-based MaxEnt harmonic grammars (Berger et al. 1996; Della Pietra et al. 1997; Goldwater & Johnson 2003; Smolensky & Legendre 2006; Wilson 2006; Hayes & Wilson 2008; White 2013). Following Becker & Gouskova (to appear), we use the term *sublexicon* to refer to each combination of a list of paradigms, a morphological operation, and a pair of MaxEnt grammars.

The first grammar (the "gatekeeper grammar") is a purely phonotactic grammar that evaluates the bases of each paradigm, while the second grammar (the "grammar proper") evaluates base–derivative alignments. On the necessity of having both types of grammars, see Becker & Gouskova (to appear). The gatekeeper grammar determines the likelihood that a given base belongs to the current sublexicon; assessing the base only, faithfulness constraints are vacuously satisfied, and only markedness constraints are operative. The grammar proper determines the likelihood of the observed derivative given its base, and thus both markedness and faithfulness constraints are potentially active. Correspondence relations (necessarily one-to-one in our implementation) are determined via application of the sublexicon's morphological operation. For example, the initial alignment of [ɹoʊz ∼ ɹoʊzɪz] gives rise to two undesirable "add [zɪ]" hypotheses, which assume a correspondence relationship between the two word-final fricatives, but thanks to the association of this paradigm with the "add [ɪz] at the right edge" sublexicon via hypothesis reduction, correspondence relations are determined correctly by the desirable alignment.

Our implementation uses a single constraint set for all grammars. The constraints are supplied by the analyst, potentially with Bayesian priors on their weights and variances. We envision future versions of the learner whose constraints are induced, potentially separately for each sublexicon (see Moore-Cantwell & Staubs 2014, Hayes & Wilson 2008).

	observed	*[+voice]#	*[+strid]#			expected
	p	$w = 2.5$	$w = 4.6$	$\mathscr{H}$	$e^{\mathscr{H}}$	p
peak	1			0	1	.48
bead	0	-1		-2.5	.08	.04
quiche	0		-1	-4.6	.01	.005
chief	.7			0	1	.48

⋮

Table 5: Training the MaxEnt gatekeeper grammar for "add [s] at right edge"

The gatekeeper grammar for a given sublexicon, e.g. the "add [s]" sublexicon, is trained by assigning the bases in this sublexicon their observed probability (which is 1 by default) and assigning all other bases of the language (if any) a probability of zero, as shown in Table 5. Each observed probability, then, is not the probability of the base *in the language*, but rather the probability of the base belonging to the "add [s]" sublexicon, e.g. [bid ∼ bidz] 'bead(s)' has a zero probability of taking an [s] plural. Each base's harmony ($\mathscr{H}$) is the weighted sum of its violations. The base's expected probability is its exponentiated harmony ($e^{\mathscr{H}}$), normalized by dividing it by the sum of all exponentiated harmonies in the sublexicon. Formally, given a sublexicon s and the i^{th} base in it b_i, the predicted probability of b_i in that sublexicon $p(b_{i,s})$ is the exponentiated harmony of $b_{i,s}$ divided by the sum of exponentiated harmonies, as shown in (1), where $w_{k,s}$ is the weight for the for k^{th} constraint in the sublexicon's gatekeeper, v_i is the vector of constraint violations for the base form b_i, and j ranges over all bases.

$$p(b_{i,s}) = \frac{e^{\mathscr{H}(b_{i,s})}}{\sum_{j=1}^{J} e^{\mathscr{H}(b_{j,s})}} = \frac{e^{\sum_{k=1}^{K} w_{k,s} v_{i,k}}}{\sum_{j=1}^{J} e^{\sum_{k=1}^{K} w_{k,s} v_{j,k}}} \tag{1}$$

The weights of the constraints are optimized by the MaxEnt module of the learner (using gradient descent) so as to minimize the sum of differences between the observed probabilities and expected probabilities. The learner can accommodate variability and noise by assigning an observed probability that is lower than one, as seen for the word "chief", here shown with a 70% observed probability for the plural [ʧifs], assuming that the remaining 30% are assigned to the plural [ʧivz]. The model's expected probability for "peaks" and "chiefs" is identical given the two constraints shown here. Additional constraints can cause the expected probability

“add z”	*[+voice]# $w = 0$	*[−strident]# $w = 0$	$\mathscr{H}$	$e^{\mathscr{H}}$
wʌg ∼ wʌgz	−1	−1	0	1

“add s”	*[+voice]# $w = 9.7$	*[−strident]# $w = 0$	$\mathscr{H}$	$e^{\mathscr{H}}$
wʌg ∼ wʌgs	−1	−1	−9.7	.00006

“add ɪz”	*[+voice]# $w = .4$	*[−strident]# $w = 11.1$	$\mathscr{H}$	$e^{\mathscr{H}}$
wʌg ∼ wʌgɪz	−1	−1	−11.5	.00001

Table 6: Candidate derivatives for the nonce [wʌg], with their harmonies.

to more closely match the observed probability.

The training increases the weights of *[+strident]# and *[+voice]#, thus correctly assigning most of the probability mass to bases that end in a voiceless non-strident. This procedure maximizes the probability of the words that belong to the “add [s]” sublexicon at the expense of the words that are observed to belong to different sublexicons, increasing the weights of the constraints that differentiate this sublexicon from the others. Our approach differs from the procedure described in Becker & Gouskova (to appear), where each sublexicon is compared to a simulation of the rich base (“salad”) that comes from the Hayes & Wilson (2008) learner, and more closely resembles the comparative phonotactics approach in Hayes (2014).

With the gatekeeper in place, the “add [s]” sublexicon expresses the limitation of [s] to voiceless non-stridents in its grammar. In contrast, the same *[+voice]# constraint will be assigned a very low weight in the “add [z]” sublexicon, since the “add [z]” sublexicon includes bases whose final segments are voiced. It will assign a low probability to bases with final voiceless sounds using a highly weighted *[−voice]# constraint. Such differences in weights between the gatekeepers allows accurate generalization to nonce words, as we discuss below.

Our choice of constraints is motivated by the distribution of plural markers in the lexicon, rather than by any substantive biases. In particular, it has been argued that languages cannot ban word-final voiceless segments (Kiparsky 2006), and thus *[−voice]# would lack support from a substantive point of view. We follow Hayes & Wilson (2008), however, who place no substantive limits on the range of permissible markedness constraints that can be induced from a lexicon (or in our case, a sublexicon).

2.3 Generalizing to novel forms

Given a nonce base, e.g. the singular [wʌg], the Sublexical Learner applies each of its sublexicons to it. Each sublexicon includes an operation, and thus generates one output candidate, to which the gatekeeper assigns a harmony score. For example, given the nonce [wʌg], the three sublexicons create [wʌgz], [wʌgs], and [wʌgɪz], as seen in Table 6. The base [wʌg] violates *[+voice]# in all three sublexicons, but this constraint has considerable weight only in the "add [s]" sublexicon. Similarly, [wʌg] violates *[−strident]# uniformly, but this violation is only heavily weighted in the "add [ɪz]" sublexicon.

Harmony scores are converted into expected probabilities of each output. These are obtained by dividing each candidate's exponentiated harmony by the sum of exponentiated harmonies; in the case of [wʌg], more than 99.9% of the probability is assigned to [wʌgz], as expected. Formally, given a derivative $d_{i,s}$ created by sublexicon s from the i^{th} base, the expected probability $p(d_{i,s})$ of that derivative is as given in (2), where s' ranges over all sublexicons. This probability can also be interpreted as the likelihood that the base b_i belongs to sublexicon s (ignoring the effect of the grammar proper).

$$p(d_{i,s}) = \frac{e^{\mathscr{H}(d_{i,s})}}{\sum_{s'=1}^{S'} e^{\mathscr{H}(d_{i,s'})}} \tag{2}$$

Naomi Feldman (p.c.) correctly points that the exponentiated harmonies given in Table 6 come from different grammars, and thus should not be compared directly without re-normalization as we do in (2), since the sum of exponentiated harmonies is different in each grammar. Another relevant factor in predicting the probabilities of derivatives is the relative size of the sublexicon that produces them, e.g. [wʌgɪz] is produced from a sublexicon that is about one tenth the size of the sublexicon that produces [wʌgz] (cf. Table 4). All else being equal, smaller sublexicons have larger constraint weights (to take away probability mass from more non-member items), which in turn favors derivatives from bigger sublexicons. Becker & Gouskova (to appear) suggest that the probability of a derivative should be multiplied by the prior probability of the sublexicon that produced it, which they take to be proportional to the size of the sublexicon, but we found that this modulation results in lower correlations with participant behavior. Robert Daland (p.c.) suggests that participants adjust the probability of a sublexicon to the experimental materials, and in a balanced experiment this would lead to a uniform probability over all sublexicons. From a practical point of view, however, we find that with neither weight re-normalization nor modulation by sublexicon size, the learner makes satisfying predictions in a variety of cases.

2.4 More English: voicing alternations

So far, the discussion of the English plural has been limited to paradigms whose derivatives do not modify any of the segments in the base. Some bases that end in [f] or [θ], however, voice the final fricative in the plural, while others keep it faithful, as seen in Table 7. This variation in existing lexical items is reflected in the treatment of nonce words, as reported by Becker et al. (2012). They gathered wellformedness ratings for final-voiced derivatives of 126 [f/θ]-final existing words and 132 nonce words. Their results suggest that voiced fricatives in the plural are preferred when following a long/tense vowel, when following a stressed vowel, and when the fricative is [f].

To model these findings in the Sublexical Learner, we augmented our original training data set of 10,000 English nouns with the 126 f/θ-final real words that Becker et al. (2012) used. Each noun was listed with its two plural candidates, including the probabilities that the participants assigned to each plural as its observed probability. The model was then trained on this combined set of 10,126 nouns and tested on the same 132 nonce words that the participants in the Becker et al. (2012) study were tested on, allowing us to assess the ability of our learner to match human behavior.

Nouns with faithful plurals, such as [bɹif ∼ bɹifs] 'brief(s)', join the other voiceless-final nouns in the "add [s] at right edge" sublexicon. When aligning a paradigm with an alternation, such as [lif ∼ livz] 'leaf/leaves', the learner identifies two changes, "add [z]" and "change [f] to [v]". To determine the generalized location of these changes, hypotheses track the possible locations for each change separately (for "add [z]": "at right edge", "after third segment", for "change [f] to [v]": "final segment", "third segment"). The correct generalization ("at right edge", "final segment") emerges victorious after hypothesis reduction.

The learner was supplied with a constraint set that allowed it to express all

	singular	plural	
alternating	lif	livz	‘leaf’
	maʊθ	maʊðz	‘mouth’
faithful	bɹif	bɹifs	‘brief’
	feɪθ	feɪθs	‘faith’

Table 7: Irregular voicing alternations for [f] and [θ]

the relevant generalizations about the distribution of voicing, e.g. a constraint that penalizes word-final [f/θ] after long vowels, etc. The complete constraint set and the rest of the simulation can be accessed at the learner’s website, `http://sublexical.phonologist.org/`.

The operations “change [f] to [v]” and “change [θ] to [ð]” are source-oriented, i.e. they mention both the source segment and the product segment (Bybee & Slobin 1982; Bybee & Moder 1983; Bybee 2001; Albright & Hayes 2003; Kapatsinski 2013a,b, see also §4). Alternatively, the use of features allows product-oriented hypotheses, i.e. those that mention the product only, e.g. “change segment to [+voice]”, allowing paradigms with [f ∼ v] alternations and [θ ∼ ð] alternations to be aggregated into a single sublexicon. As Figure 3 shows, combining [f] and [θ] words in the same sublexicon offers a substantial improvement to the prediction of the participants’ behavior, raising Spearman’s ρ from .17 to .51.[2] The correlation in the right hand panel of Figure 3 is not just quantitatively stronger; it also shows the effect of the expected predictors of alternation qualitatively. Just like speakers do, the model assigns the highest probability to voicing of [f] following a long, stressed vowel, and the lowest probability of voicing to [θ] following an unstressed vowel.

The learner creates both source-oriented and product-oriented hypotheses, and hypothesis reduction favors the product-oriented, feature-based generalizations in this case thanks to their better coverage. See, however, Becker & Gouskova (to appear) for a source-oriented generalization in Russian that cannot be expressed as a product-oriented one.

While the [f/θ]-final nouns are the largest group of unfaithful plurals in English, they do not exhaust the list. There are also various Latinate plurals (e.g.

[2] These are the correlations between the model’s predictions and the aggregate participant rating for each of the 132 nonce words in Becker et al. (2012). The correlation coefficients we report are the mean ρ obtained from a Spearman’s rank correlation test, repeated with jitter 10,000 times, using R’s *cor.test* function.

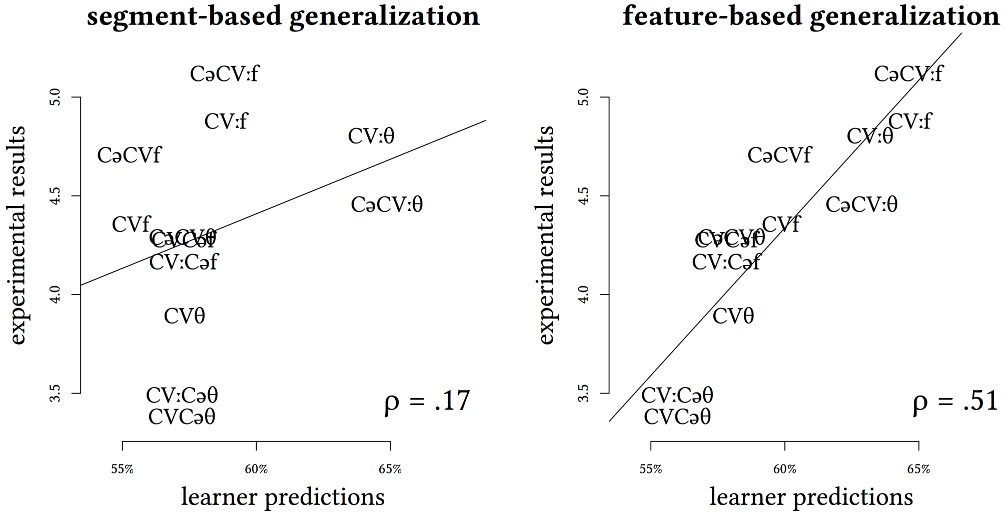

Figure 3: The learner's ability to predict the acceptability of voicing in nonce words, learning with separate treatment of [f ∼ v] and [θ ∼ ð] (left) and with unified treatment of the two alternations (right).

radius ∼ radii), unchanged plurals (e.g. deer ∼ deer) and vowel changing plurals (e.g. goose ∼ geese). Since we have no evidence about the productivity of these patterns, and our intuition is that their productivity is rather low, we chose to eliminate them from consideration, which we did by imposing a minimal size threshold of 10 paradigms per sublexicon. This threshold prevents sublexicons with fewer than 10 associated paradigms from being applied to test items. The threshold value was set arbitrarily in this case, but as we will see in §4, there is reason to think that imposing a minimal size restriction on sublexicons may lead to superior predictions of nonce task results by preventing overfitting of the training data. The purpose of making sublexicons is to extend productive patterns from the lexicon to nonce words; extending unproductive patterns and assigning some of the probability mass to them takes away probability mass from the forms that speakers actually prefer.

2.5 Comparison to representational approaches

Our model contrasts rather sharply with most work in generative phonology, as it does not posit an underlying representation for the plural morpheme, nor for any of the nouns. Rather, the entire burden of the analysis is shifted to the set of grammars, with the number of grammars and their coverage of the lexicon determined according to the algorithm in §2.1; there is no worry about an unprincipled proliferation of grammars (cf. Inkelas & Zoll 2007). Additionally, compared to most work in constraint-based phonology, we severely limit the role of Gen, the candidate generation function (Prince & Smolensky 1993/2004). The number of candidates is equal to the number of sublexicons, and there is no further manipulation of the output candidates.

In representational approaches, the (underlying) representation of the suffix is chosen to be /z/, and the mapping from /z/ to the unfaithful [s] and [ɪz] relies on two general facts about English: a prohibition on voicing disagreement in final obstruent clusters, and a prohibition on final adjacent stridents. Our model produces plural morphology correctly without any access to these language-wide restrictions, showing that they are not logically necessary.

Due to our learner's indifference to language-wide generalizations, it learns allomorphy with equal success whether the allomorphs optimize anything in the output or not. While the English plural can be thought to optimize voicing and stridency, non-optimizing affixation is well-attested in a wide variety of languages (Bobaljik 2000; Paster 2006; Bye 2007; Embick 2010; Nevins 2011; Gouskova & Newlin-Łukowicz 2014). Indeed, non-optimizing phonological restrictions are easy to find in English as well, e.g. the comparative *–er*, which attaches to trochaic bases only (e.g. small-er, happy-er, *morose-er). Tellingly, the agentive *–er* attaches to any base indiscriminately, showing that the language-wide phonology places no restrictions on such affixation. Given an appropriate constraint set, our learner treats optimizing and non-optimizing affixation with a similar mechanism, arguably like humans do. It may be the case that language-wide restrictions *facilitate* morphophonological learning; if so, our learner's disinterest in them would be a departure from human behavior. See Moore-Cantwell & Staubs (2014) for a model that is somewhat similar to ours, but with the express purpose of reaching underlying representations that allow language-wide restrictions to play a role.

The constraints that are active in our analysis are those whose violations serve to separate one sublexicon from another. Thus, *[+voice]# is violated by the base in all "add [z]" words and some "add [ɪz]" words, but no "add [s]" words, and thus it is weighted highly in the "add [s]" sublexicon. In contrast, *ComplexOnset is violated approximately equally in all sublexicons, and therefore its weight stays close to zero in all sublexicons.

Constraints that are never violated in any word of the language, and therefore never violated in any sublexicon, such as OCP(strident), will not change their initial weight. In a traditional Optimality Theoretic account, OCP(strident) is central to the analysis, with the ranking OCP(strident) $\gg$ Dep driving epenthesis in /kɪs + z/ $\rightarrow$ [kɪsɪz] 'kiss'. In our sublexical analysis, no observed forms violate OCP(strident), and if its initial weight is zero, this weight will not change, and the constraint will remain inert.

However, if OCP(strident) receives a high initial weight — plausibly, from a phonotactic grammar of the language as whole — and if it is entered into the sublexical analysis of the plural with this weight, then its weight will remain high in all sublexicons. Similarly, Agree(voice) is not violated in any forms, and its weight will remain unchanged in the sublexical analysis. These constraints are insufficient for a complete sublexical analysis, however, e.g. *[+voice]# is needed to block the addition of [s] to sonorant-final nouns. The sublexical analysis, then, is entirely compatible with, and orthogonal to, a wider view of the language.

Another way to integrate language-wide phonotactics would be to fit a language-wide MaxEnt phonotactic grammar (e.g. using the UCLA Phonotactic Learner, Hayes & Wilson 2008) in addition to our sublexical analysis, and then combine the two analyses by adding their harmony scores, or equivalently, multiplying their predicted probabilities.

In addition to the sublexical analysis of the plural and the general phonotactic grammar of the language, it is likely that a full representation of speaker knowledge requires other grammatical analyses at varying levels of generality. For example, English does not allow final [fs] in singular nouns, nor in verb roots, but the same cluster is allowed in plural nouns or third person singular verbs (e.g. [kʌfs] 'cuffs'). Since *[fs]# is never violated by singular nouns, its weight will remain at zero in all gatekeeper grammars, and its weight will likely be very low in the language-wide phonotactic grammar as well. Its effect will only be captured by a different grammar, one that separates singular nouns from plural nouns. Perhaps more generally, then, the probability of a paradigm should be the combination (by addition or multiplication) of all the grammars that the speaker uses in their analysis of the language.

To summarize, we have presented a general-purpose mechanism for learning morphophonology from two-word paradigms, exemplified using the English plural suffix. Our learner first creates a large space of of hypotheses which then undergoes reduction to the smallest set able to account for all the training data. Once this final set of hypotheses is identified, two MaxEnt grammars are fitted to each one, thus learning for each a productive morphological operation and any phonological restrictions on its application. The procedure correctly generalizes regular (exception-free) distributions as well as gradient (or noisy or variable) distributions.

3 Beyond simple suffixation

The Sublexical Learner's ability to identify the content and position of affixes was exemplified in §2 above with a rather simple case of suffixation. In this section, we demonstrate its coverage of a broader range of morphological exponents, and in particular the kinds of morphophonology that are inaccessible to the Minimal Generalization Learner. We start in §3.1 with cases of multiple exponence in two dialects of Berber, wherein one change is left-oriented and the other is right-oriented. Metathesis is treated with an example from Rotuman in §3.2. An infixation pattern from Tagalog in §3.3 motivates an elaboration of the hypothesis consumption algorithm. The need for suprasegmental loci is shown with an example from the English past tense in §3.4. We summarize and suggest future directions in §3.5.

Our implementation includes all of the cases discussed in this section. The simulations may be found at `http://sublexical.phonologist.org/`.

3.1 Locating multiple exponence in Berber

When given a morphological exponent that involves multiple changes, our learner creates hypotheses about the location of each change individually, as seen briefly in §2.4. We demonstrate this ability here with two cases of Berber circumfixation. The first is the Imdlawn dialect, where the feminine is marked with a [t- -t] circumfix, e.g. [funas ∼ tfunast] 'bull/cow'. When more than one change is identified, the position of each change is stated in left-oriented and right-oriented terms, and the statements are combined into hypotheses.

In the case of [funas ∼ tfunast], the initial [t] gives rise to the left-oriented "add [t] before the first segment" and right-oriented "add [t] before the prepreantepenultimate segment"; the final [t] gives rise to left-oriented "add [t]

	masculine	feminine	
a.	akli	t–akli–t	‘slave’
b.	amawat	t–amawat	‘conductor’
c.	anabrak	t–anabrak	‘digger’
d.	analmad	t–analmat	‘pupil, apprentice’
e.	amaksad^{ʕ}	t–amaksat^{ʕ}	‘coward’
f.	amənhug	t–amənhuk	‘lacking sense’
g.	amahaʁ	t–amahaq	‘Tuareg person’

Table 8: Tuareg Berber circumfix with devoicing/stopping at the right edge

after the fifth segment” and the right-oriented “add [t] after the last segment”. These change are paired to yield four hypotheses, one of which is the desirable “add [t] before the first segment and add [t] after the last segment”. Given a small dataset, taken from Elmedlaoui (1995), the learner correctly consumes all other hypotheses.

A more complex learning problem can be seen in the feminine forms of the Tuareg dialect, shown in Table 8 (Elmedlaoui 1995, citing Prasse 1972). The left side of the circumfix is uniformly [t]. The right side appears as [t] after vowel-final bases (Table 8a), or as no change after voiceless plosives (Table 8b-c). If the base ends in a voiced obstruent, it is devoiced and stopped, potentially vacuously (Table 8d-g).

During hypothesis creation, the learner produces “add [t] before the first segment and add [t] after the last segment” (Table 8a), “add [t] before the first segment (and no other change)” (Table 8b-c), “add [t] before the first segment and devoice final segment” (Table 8d-f), and “add [t] before the first segment and devoice and stop final segment” (Table 8g), in addition to many others. The last three of these can in fact be consumed by “add [t] before the first segment and devoice and stop final segment”, assuming that product-oriented generalizations also allow vacuous application. The operation states that the final segment of the base must become [−continuant −voice], without regard to its original specification for these two features.

The consuming hypothesis is created by the combined devoicing and stopping of [ʁ]-final stems, but its coverage includes all of the plosive-final stems as well, and is thus allowed to consume the devoicing-only and no-change hypotheses. This is consistent with the view we offered in §2, that the consuming hypothesis

	complete	incomplete	
a.	hosa	hoas	‘flower’
	pure	puer	‘to rule’
b.	hoti	høt	‘to embark’
	futi	fyt	‘to pull’

Table 9: Rotuman metathesis

must have greater coverage than the consumed hypothesis.

3.2 Metathesis in Rotuman

The Rotuman “incomplete phase” (McCarthy 2000, citing Churchward 1940, see also Blenkiron & Alderete 2015) is expressed by metathesis of a verb-final consonant and vowel (Table 9a). When the stem’s final vowel is front, and not lower than its penultimate vowel, the stem’s penultimate vowel is fronted, and the final vowel is deleted (Table 9b). We do not attempt to cover the full range of data, which has been discussed extensively in McCarthy (2000) and references within; we simply wish to show how our learner analyzes metathesis.

When trained on a small Rotuman dataset similar to the one in Table 9, only two sublexicons emerge after hypothesis consumption: “metathesize the last two segments” and “front the antepenultimate segment and delete the final segment”, with a product-oriented operation unifying the fronting of the two back vowels. The two sublexicons are associated with gatekeeper grammars that are sensitive to the backness and height of the last two vowels of the stem, and thus the two patterns are generalized correctly.

The learner uses a symbolic, segment-blind metathesis operation, which is therefore predicted to apply productively to any pair of segments, going beyond the training space of the observed vowels and consoanants (Berent et al. 2012).

The sublexical analysis uses metathesis when it is observed in the surface forms (Table 9a), but not when masked by deletion (Table 9b). It thus differs from the unified metathesis and fusion analysis of McCarthy (2000); yet one would be hard pressed to distinguish the two analyses empirically. In addition to its empirical adequacy, our sublexical analysis has the advantage of being demonstrably learnable directly from the surface forms that are available to the learner.

		base	derivative	
a.	#V	abot	in–abot	‘attain’
b.	#CV	bago	b–in–ago	‘big’
c.	#CCV	pɾoblema	p–in–ɾoblema pɾ–in–oblema	‘problem’

Table 10: Tagalog [in]-infixation (Zuraw 2007)

3.3 Multiple consumption in Tagalog

A challenging case of hypothesis consumption comes from the Tagalog infix [in], described and analyzed in Zuraw (2007). When the stem begins with a single consonant, [in] is invariably infixed after this initial consonant (Table 10b). When the stem begins with a consonant cluster, [in] can be inserted either after the first consonant or after the cluster (Table 10c). As for words that begin with a vowel (Table 10a), the affix appears initially. One could analyze vowel-initial words as involving an initial glottal stop, as done in Halle (2001), to unify their treatment with the treatment of single-consonant initial words and thus reduce the complexity of the description; we deliberately choose the more challenging characterization, to ensure that our learner can handle it.

The learner’s goal is to discover that the infix is left-oriented, i.e., that its position is computed relative to the left edge of the word. The learner generates both left-oriented and right-oriented hypotheses, as described in §2 above. In the Tagalog case, however, the undesirable right-oriented hypotheses cannot be consumed by any single left-oriented hypothesis; the situation is schematized in Table 11, with 3-, 4- and 5-segment words, where “x” stands for any segment. The hypothesis “add [in] before the penultimate segment” only applies to three- and four-segment words, and thus should not be generalized to longer words, but there is no single left-oriented hypothesis that can consume it; it can only be consumed jointly by “add [in] after the first segment” together with “add [in] after the second segment”.

The Tagalog case shows that hypotheses must be given the chance to jointly consume other hypotheses, or more generally, that the learner should be searching for the smallest set of hypotheses that can jointly account for all of the data. Our hypothesis reduction algorithm must be amended from the original in Table 3 to the more powerful formulation in Table 12.

left-oriented		right-oriented
before first segment	in x x x	before antepenult
before first segment	in x x x x	before preantepenult
before first segment	in x x x x x	before prepreantepenult
after first segment	x in x x	**before penult**
after first segment	x in x x x	before antepenult
after first segment	x in x x x x	before preantepenult
after second segment	x x in x	before last
after second segment	x x in x x	**before penult**
after second segment	x x in x x x	before antepenult

Table 11: Desirable hypotheses are not supersets of undesirable ones

Hypothesis H_L is consumed by set of hypotheses $\{H_{W1}, H_{W2}, ...\}$ iff
$\{H_{W1}, H_{W2}, ...\}$ and H_L are *idempotent* on H_L's bases
and $\{H_{W1}, H_{W2}, ...\}$ cover *more lexical items* than H_L.

Table 12: Hypothesis reduction: multiple consumption (final, cf. Table 3)

The algorithm in Table 12 is currently implemented by searching the power set of hypotheses in order of increasing cardinality. The algorithm starts by allowing each hypothesis to try and consume the entire lexicon; if this fails, all pairs of hypotheses are tried out, then triplets, etc., until the entire lexicon is covered by the smallest possible number of hypotheses. We use the term *single consumption* for the cases described in §2, where one hypothesis can consume another, and *multiple consumption* for cases like Tagalog, where multiple hypotheses are needed to conjointly consume a undesirable hypothesis.

Our proposed technique for performing multiple consumption may be accurately characterized as a bruce-force exhaustive search; we content ourselves with finding a solution that works at this point, with the hope for a more expedient search algorithm in the future. Even with the current procedure, however, the learner is acceptably efficient for all of the data sets described in this paper, and does not try the linguist's patience too much.

Even in the Tagalog data, where single consumption is insufficient, single consumption is still useful in greatly reducing the number of initial hypotheses. We therefore speed up the search by first performing single consumption, then performing multiple consumption on the smaller set of remaining hypotheses.

Multiple consumption is needed whenever the location of a morphological exponent is non-uniform. Most such cases are traditionally analyzed using techniques such as floating features (see Wolf 2007). One such example would be Chaha's expression of the 3SG.MASC object pronoun as labialization that can impact the final segment of a verb, e.g. [dænæg ~ dænæg^w] 'hit' or its antepenultimate segment, e.g. [kæfæt ~ kæf^{w}æt] 'open' (McCarthy 1983), depending on the place of articulation of the relevant segments.

3.4 Suprasegmental loci in English

We offer a comprehensive analysis of the English past tense in §4. Here, we only discuss two of the many patterns: the stem vowel change to [ʌ] (e.g. *win* ~ *won*) and the stem vowel change to [oʊ] (e.g. *ride* ~ *rode*).

The vowel [ʌ] originates from [ɪ] in *dig* ~ *dug* and from [i] in *sneak* ~ *snuck*. A product-oriented generalization such as "change segment to [ʌ]" will allow the mappings to [ʌ] from both [ɪ] and [i] to be included in the same sublexicon. For *dig* and *sneak*, the change is in the penultimate segment, yet the change occurs in the antepenultimate segment of *drink* ~ *drunk*. Clearly, the actual location is "last vowel", or rather, "last nucleus", not "penultimate segment".

To allow the learner to discover the generalization, we added "last nucleus" as a possible location of change. From there, this location's wider coverage allows hypothesis reduction to consume "penultimate segment" and "antpenultimate segment", and home in on "change last nucleus to [ʌ]" as the correct operation.

The change to [oʊ] should similarly target the last nucleus. Yet in the existing verbs of English, this change always happens before simple codas (*ride* ~ *rode*, *weave* ~ *wove*, *break* ~ *broke*, *choose* ~ *chose*, etc). For the learner, "change last nucleus to [oʊ]" and "change penultimate segment to [oʊ]" have the exact same coverage. While the two generalizations make the same predictions for a verb with a final simple coda, the generalizations differ for complex coda: a nonce [misk] would receive the past tense [moʊsk] from the former generalization and [mioʊk] from the latter. Since the latter is clearly worse, we augmented the learner with a bias in favor of supra-segmental hypotheses in cases of equal coverage.

This same issue emerged in the sublexical analysis of Russian yer deletion (Becker & Gouskova to appear), wherein mid vowel deletion always happens before a simple coda, e.g. [rot ~ rta] 'mouth NOM/GEN'. Here, "delete last nucleus" and "delete penultimate segment" have the same coverage, and the former is chosen thanks to the bias in favor of supra-segmental hypotheses.

3.5 Summary and future directions

At present, our learner's capabilities cover an encouragingly wide range of morphological relations, including many cases of infixation, circumfixation, subtractive morphology, internal stem changes, etc. Yet some morphophonology is still beyond its reach. For example, Ulwa's infixation pattern (Table 2 in §2) requires the location "after first foot", which is currently not a change locus that the learner is able to express.

The treatment of stress and tone will likely require further operations and/or change loci, such as "destress all nuclei", "make all nuclei High", and others. Special treatment will also be required by reduplication patterns, allowing the learner to notice similarities between the stem and the segments that are added to the derivative.

Finding the complete range of needed operations and loci is a non-trivial task, since, as we have demonstrated, the generalizations that are formed by our Sublexical Learner are different from those hypothesized by generative linguists, a difference that is mainly due to our focus on unsupervised learning of the morphological exponents. What may seem daunting to the generative linguist may not seem daunting to the Sublexical Learner, and vice versa. A case in point is the Arabic plural, which is famously expressed in a wide variety of ways (see McCarthy & Prince 1990; Dawdy-Hesterberg 2014, a.o.), and whose sublexical analysis would therefore require a large number of sublexicons. A preliminary look suggests, however, that the analysis may be fairly manageable even with only the existing machinery. For example, many CVCC nouns take the ʔaCCaːC plural, e.g. [milk ∼ ʔamlaːk] 'property', [qutˤb ∼ ʔaqtˤaːb] 'pole', [θawb ∼ ʔaθwaːb] 'garment'. These can share the sublexicon "add [ʔa] before the first segment, delete second segment, add [aː] after third segment", and similarly in other cases.

More generally, we recognize that the expression of morphological relations is likely not generally computed by counting segments, e.g. "after the fourth segment" is likely not a locus that speakers use. McCarthy & Prince (1998), for example, point to the prosodic hierarchy as a universal theory of change loci, although the range of phenomena they examine is different from ours. Relatedly, McCarthy (2003) proposes that distance from the edge is assessed categorically, not counted using any units. Our ultimate goal is to learn all known types of morphology and generalize them in the same way that native speakers do.

	content	position	distribution
a.	[t]	right edge	after voiceless segments, except [t]
b.	[d]	right edge	after voiced segments, except [d]
c.	[ɪd]	right edge	after the alveolar plosives [t, d]

Table 13: The English regular past

4 The English past tense

The literature on the English past tense is vast (for a recent overview, see Seidenberg & Plaut 2014 and references within), and it overwhelmingly uses analogical models that require the analyst's supervision to align bases and derivatives. This is true as well for dual-route models (Prasada & Pinker 1993 et seq.), which contain an analogical component. Since our focus is on learning the base-derivative alignments without supervision, and since we have amply demonstrated that this alignment is a non-trivial task, we leave analogical models and their known strengths aside. Our approach follows instead in the footsteps of Albright & Hayes's (2003) analysis of the English past tense, which succeeds without any supervision in creating mappings from bases to derivatives.

Our study of the English past tense relies on the same training set of 4253 verbs used by Albright & Hayes (2003). In this set, 95% of the verbs express the past tense using a [d], [t], or [ɪd] suffix, with the familiar distribution in Table 13. The Sublexical Learner easily finds three sublexicons for these verbs.

Once trained on Albright & Hayes's 4253 verbs, the model was tested on the same 58 nonce verbs that they used in their experiments and modeling. When a productivity threshold of 100 is imposed (i.e. sublexicons are required to contain at least 100 paradigms), the model finds the three expected sublexicons from Table 13, and thus generates regular past tense forms 100% of the time. The constraint set we used included various right-edge restrictions, e.g. *[+voice]#, *[t,d]#, *[ɪ]C_0#, among others. The full constraint set can be found at the learner's website, `http://sublexical.phonologist.org/`. With these constraints, the distributional restrictions in Table 13 were learned as well. Since the experimental participants usually preferred regular forms both in their productions and in their ratings, the model's predictions achieve a rather strong correlation with the participants' preferences ($\rho = .70$).[3]

[3] Here and throughout, we report the correlation between the model's predictions and Albright

	Sublexicon	Coverage	
1.	Add [d] at right edge	2104	ʤɑsəl, spɑɹ, vɪʒuəlaɪz, ɛkoʊ, ...
2.	Add [ɪd] at right edge	1146	hænd, koʊɹt, sʌfəkeɪt, sʌbstɪtut, ...
3.	Add [t] at right edge	791	ɪŋkɹis, tɹɛspæs, mʌf, gɑsɪp, pɚʧ, ...
4.	Make last nucleus [oʊ]	30	wiv, stɹaɪv, fɹiz, stil, teɪɹ, ɪntɚwiv, ...
5.	No change	29	lɛt, ʌpsɛt, bit, ɹisɛt, wɛt, kwɪt, ...
6.	Make last nucleus [ʌ]	20	spɹɪŋ, dɪg, ɹɪŋ, wɪn, snik, slɪŋ, ...
7.	Make last nucleus [ɛ]	18	fɔl, fid, hoʊld, bɹɛstfid, wɪθhoʊld, ...
8.	Make last nucleus [æ]	15	ɹʌn, sɪt, ɹɪŋ, bɪgɪn, spɹɪŋ, aʊtɹʌn, ...
9.	Add [t] at right edge, make last nucleus [ɛ]	12	slip, kɹip, nil, fil, dil, kip, swip, ...
10.	Make last nucleus [u]	11	θɹoʊ, bloʊ, flaɪ, dɹɔ, aʊtgɹoʊ, ...
11.	Make last nucleus [eɪ]	10	it, fɚgɪv, kʌm, laɪ, oʊvɚit, fɚbɪd, ...
12.	Make last nucleus [ʊ]	8	mɪsteɪk, ʃeɪk, teɪk, pɑɹteɪk, ...

Table 14: English past tense sublexicons with a threshhold of $\geq$8

To cause the model to produce irregular past tense forms, as the participants did, the productivity threshold was lowered. The lowest threshold we tested was 8, which gives rise to 12 sublexicons, listed in Table 14. With these twelve sublexicons in place, the learner created twelve past tense forms for each of the 58 nonce verbs, e.g. for the present tense [baɪz] the learner created the past tense forms [baɪzd], [baɪzɪd], [baɪzt], [boʊz], [baɪz], [bʌz], [bɛz], etc., each with its predicted probability.

The probabilities that the model assigned were compared to the mean probability of production by Albright & Hayes's (2003) participants. If a past tense from the experiment was not produced by our learner, it was added with a probability of zero; this was the case for only a few forms, e.g. [dæpt] as a past tense for [deɪp]. With its default settings, the sublexical learner reached a correlation of $\rho = .72$ with the participants' productions. A visual inspection suggested that high constraint weights in the smaller sublexicons left the regular forms with too much of the probability mass relative to the participants' productions. To limit the ability of constraint weights to get too large and thus overfit the smaller patterns in the data, the variance (σ) of the Bayesian prior was reduced from the default 100000 to 10, resulting in more of the probability mass being given to irregular

& Hayes's (2003) participants' productions (i.e. the probability that a given form was produced by the speakers). In all cases, we report the mean ρ obtained from a Spearman's rank correlation test, repeated with jitter 10,000 times, using R's *cor.test* function.

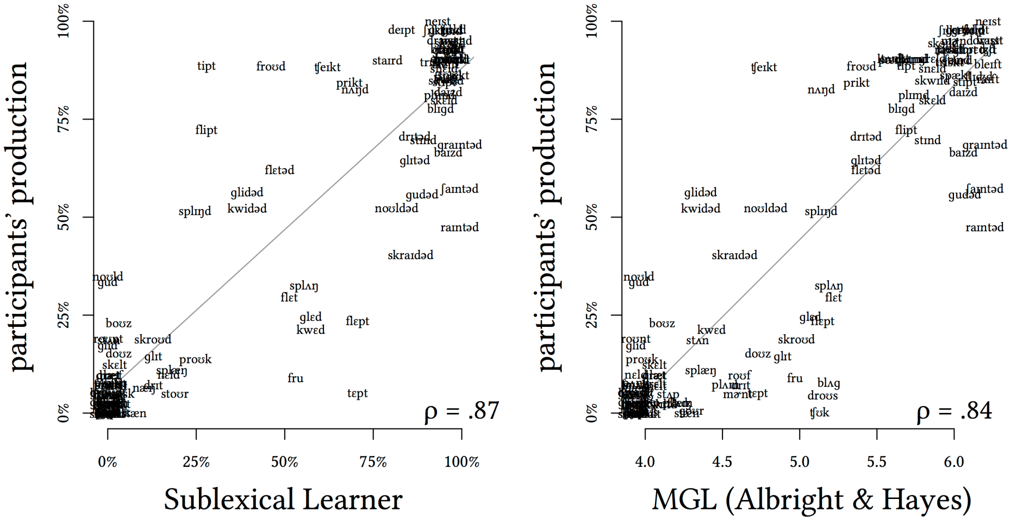

Figure 4: Participants' productions of English past tense forms, as predicted by our Sublexical Learner and the Minimal Generalization Learner.

forms. In particular, more of the probability mass was given to forms that change the last stem vowel to [oʊ] or [ʌ] (sublexicons 4 and 6). This improved the learner's predictions to $\rho = .79$. The addition of this type of Bayesian prior on constraint weights follows Wilson (2006), among many others (see also White 2014).

The goodness of [oʊ] and [ʌ] as past tense vowels, regardless of the last stem vowel in the present, is captured by our model's product-oriented operations. The product-oriented operation "change last nucleus to [oʊ]" unified the four different vowels [aɪ, eɪ, i, u] that change to [oʊ], while these remain four different operations in Albright & Hayes (2003) and in Moore-Cantwell & Staubs (2014). We thus supply a direct answer to the call for product-oriented generalizations in Albright & Hayes (2003; §5.2).

After reducing σ, our inspection suggested that forms created by the smaller sublexicons (e.g. "make last nucleus [ʊ]") were getting more than their fair share of the probability mass. To explore this possibility, starting from our original threshold of 100, we gradually decreased the productivity threshold, measuring the correlation with the participants' productions as the number of sublexicons increases. The correlation starts at $\rho = .72$ (as noted above) with three sublexicons, and increases rapidly with the addition of sublexicons 4, 5, 6, and 7, when it reaches $\rho = .86$. It then continues to climb very slowly with the addition of sublexicons 8, 9, and 10, reaching a maximum of $\rho = .87$. This is our best model, whose predictions are graphed in the left panel of Figure 4.

Further decrease in the productivity threshold, adding sublexicon 11 and then sublexicon 12, reduced the correlation back to $\rho = .86$ and then $\rho = .84$. The creation of sublexicons 11–12 undeservedly assigned some of probability mass to operations that participants hardly deployed at all, leading to deteriorated predictions. Our best model, then, uses 4176 verbs to make nonce word predictions, leaving 77 verbs (2% of the total) as unproductive items that need to be memorized but not generalized. We find it reassuring that the best number of sublexicons is rather moderate, because our learner's computational load increases dramatically with the number of sublexicons; in particular, the step of multiple hypothesis reduction (§3.3) searches the hypothesis power set, and thus grows exponentially with the number of hypotheses.

In comparison, the Minimal Generalization Learner (MGL) generalizes from patterns of any size, and the computational load of a pattern is proportional to its attestedness; the more lexical items are involved, the more rules are created. The MGL's predictions are plotted in the right panel of Figure 4, with a correlation that is almost as high as our learner's ($\rho = .84$). An ANOVA nested model comparison of the two learners did not indicate a winner, but rather showed that each learner contributes significantly to predicting the participants' productions beyond the ability of the other.[4]

Another analysis of the English past tense that is similar in spirit to ours is offered in Moore-Cantwell & Staubs (2014). Their learner is based on finding classes of paradigms that are derived by identical operations, like ours, but their operations are not necessarily surface-true. For example, their learner unified "add [d]", "add [t]", and "add [ɪd]" to a single "add /d/" operation. They do so in order to give a role to language-wide phonotactic constraints such as *DD# (no two alveolar plosives word-finally). As we argue in §2, however, an approach that does not rely on language-wide phonotactics generalizes straightforwardly to cases where the morphophonology does not respond to general constraints, and often seems non-optimizing. Such non-optimizing morphology is very common, and it is not clear that language learners disfavor it in any way.

We conclude that our learner compares favorably to the Minimal Generalization learner, and further improves on it by forming the product-oriented hypothe-

[4] We compared the models by converting their predictions and the participants' productions to rank scores, then fitting linear models and comparing with an a ANOVA using the *lm* function in R. Both learners contribute significantly to the superset model, with our model's contribution being larger ($\chi^2(1) = 41.4$ vs. $\chi^2(1) = 20.3$, both highly significant).

ses that Albright & Hayes's (2003) allude to. With our hand-made constraints, the learner generates predictions that correlate very strongly with the participants' responses, and we hope that a future version with algorithmically generated constraints will further improve on this result.

5 Conclusions

We have presented an implementation of the Sublexical Learner, which takes two-word paradigms as inputs and learns the mapping from one set of forms to another by extracting operations and restrictions on their distributions. It then uses this knowledge to generate derivatives for novel words, each with its predicted probability.

Our learner shares several central assumptions with the Minimal Generalization Learner (MGL, Albright & Hayes 2002, 2003, 2006). First, we assume that morphophonological learning starts once the speaker has arranged pairs of words into paradigms, e.g. once they have decided that [dʌk] 'duck' and [dʌks] 'ducks' are related, they can discover an "add [s]" operation. Second, we work based on the surface forms of the paradigm members, without a search for any underlying forms. No hidden structure is attributed to lexical items beyond their sublexical membership. Third, we assume that phonological effects in the lexicon are extended to novel forms via competing generalizations; for the MGL, rules compete within a single grammar, while for the Sublexical Learner the sublexicons compete, as well as the constraints within each sublexicon. Fourth, we assume that morphophonological learning proceeds largely independently of the overall phonology of the language. Both learners, however, allow the analyst to include the influence of grammar-wide constraints. Due to these similarities, much of the results and insights in Albright (2002a,b, 2006, 2008a,b) apply rather straightforwardly to our learner as well.

The main differences between our learner and the MGL largely stems from our separation of operations from their loci and conditioning environments. This difference creates a number of distinct advantages. First, in cases of multiple changes, it allows us to track the locus of each change individually (§3.1). Second, it allows us to learn non-local conditioning factors, e.g. the selection of the English comparative suffix *–er* by trochaic bases, or the selection of the [n] allomorph of the Berber reciprocal prefix when the base contains a labial in any position

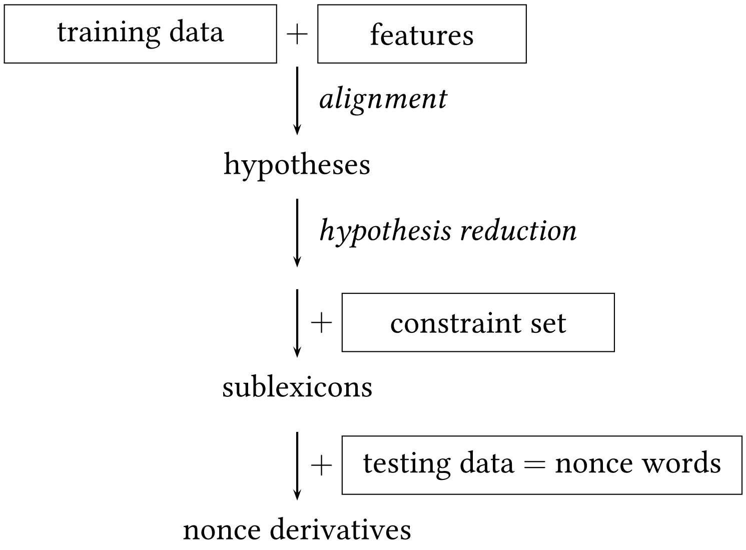

Figure 5: The Sublexical Learner

(Elmedlaoui 1995). Third, it allows us to explore hypotheses about change loci, including variable infixation loci (§3.3) and supra-segmental loci (§3.4) using our mechanism of hypothesis reduction. Additionally, our use of MaxEnt grammars allows us to learn from variable/noisy data using the well-understood class of Maximum Entropy models; this aspect of the model allows it make probabilistic predictions with a solid mathematical basis, improving on the confidence and reliability scores produced by the MGL.

The operation of the learner is schematized in Figure 5, where the analyst's input is in framed boxes. The training data is a list of two-word paradigms, and the testing data is a list of bases. The feature chart is used to improve the alignment of bases and derivatives (see §2.1) and to evaluate constraint violation. The analyst's supervision is required in the form of a constraint set, hopefully to be replaced in the future by a constraint induction mechanism (e.g. Hayes & Wilson 2008; Moore-Cantwell & Staubs 2014).

We have contrasted our model with approaches that attempt to learn a single morpheme-blind grammar (Tesar & Smolensky 1998, 2000; Tesar et al. 2003; Riggle 2006; Jarosz 2006; Rasin & Katzir 2013; among others), and have suggested that this assumption likely prevents large-scale learning from realistic data, as it runs counter to a large body of theoretical work in phonology, in particular the literature on constraint indexation (Pater 2000, 2006, 2008; among many others).

We accommodate non-uniformity in the lexicon by adopting multiple grammars. Others have attempted to keep a single grammar and enrich it with morpheme-specific constraints, e.g. Pater et al. (2012) and Moore-Cantwell & Staubs (2014), in some cases including constraints that enforce operations, as in Moore-Cantwell & Staubs (2014); Hayes (1999a,b); Albright & Hayes (2006). In particular, the learner in Moore-Cantwell & Staubs (2014) shares many characteristics with our learner, including the bottom-up aggregation of words into sublexicons ("bundles" in their terms), cf. top-down approaches that start with a single lexicon that is subsequently partitioned (Pater 2006, 2008; Becker 2009).

In approaches based on Minimal Description Length (MDL, e.g. Goldsmith 2001; Rasin & Katzir 2013), the training data consist of individual words, with a learner that connects bases with derivatives through their shared segments. While the implementations that are known to us rely on a single morpheme-blind grammar, one can imagine that an MDL-based approach could be combined with a sublexical approach. We hope to explore such possibilities in the future.

Finally, our learner is designed to learn rather fine-grained generalizations, examining the mapping from one morphological category to another (e.g. singular to plural), ignoring other mappings (e.g. plural to singular, nominative to genitive, etc.) and any language-wide generalizations. There is ample reason to believe, however, that this category-specific knowledge only partially describes the language speaker's knowledge, which likely includes phonological generalizations about wider or overlapping categories, such as animate nouns, native verbs, etc. Speakers seem to transfer knowledge between lexicons, e.g. modulate the application of an operation based on language-wide generalizations, or apply an operation beyond its expected scope (Bybee & Slobin 1982; Kapatsinski 2013a,b; a.o). Additionally, knowledge from one suffix may be used in the treatment of other suffixes, e.g. all Russian nouns that undergo vowel deletion before the genitive suffix [a] (e.g. [rot ∼ rt-a] 'mouth.NOM/GEN') also undergo deletion before the other case suffixes. It seems likely that speakers generalize these regularities between suffixes, which our current implementation does not capture. Of course, uniform behavior of affixes is far from universal: an English example includes voicing of [f] in plurals and denominal verbs, where e.g. the [f] of [bəlif] 'belief' voices in the verb [tə bəliv] but not in the plural [bəlifs], while the [f] of [naɪf] 'knife' voices in the plural but not in the verb (Becker et al. 2012). Our current implementation is compatible with the existence of such broader generalizations, and we hope to explore these topics in future work. One possibility is that the well-formedness of a derivative is a combination (addition or multiplication) of all the grammars that assign a probability to the derivative, its base, or their relation.

References

Albright, Adam (2002a). *The Identification of Bases in Morphological Paradigms.* Ph.D. dissertation, UCLA.

Albright, Adam (2002b). The lexical bases of morphological well-formedness. In sabrina bendjaballah, w. u. dressler, o. e. pfeiffer & m. voeikova (eds.) *Morphology 2000: Selected papers from the 9th Morphology Meeting, Vienna, 24-28 February 2000*, Benjamins. 1–8.

Albright, Adam (2006). Inflectional paradigms have bases too: Arguments from Yiddish. Ms. MIT.

Albright, Adam (2008a). A Restricted Model of UR Discovery: Evidence from Lakhota. Ms. MIT.

Albright, Adam (2008b). Explaining universal tendencies and language particulars in analogical change. In Jeff Good (ed.) *Language Universals and Language Change*, Oxford University Press. 36 pp.

Albright, Adam & Bruce Hayes (2002). Modeling English past tense intuitions with minimal generalization. In Michael Maxwell (ed.) *Proceedings of the sixth meeting of the ACL special interest group in computational phonology*. Philadelphia: ACL, 58–69.

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

Albright, Adam & Bruce Hayes (2006). Modeling productivity with the gradual learning algorithm: The problem of accidentally exceptionless generalizations. In Gisbert Fanselow, Caroline Féry, Matthias Schlesewsky & Ralf Vogel (eds.) *Gradience in Grammar*, Oxford University Press. 185–204.

Anttila, Arto (2002). Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory* **20**. 1–42.

Becker, Michael (2009). *Phonological Trends in the Lexicon: The Role of Constraints.* Ph.D. dissertation, University of Massachusetts Amherst.

Becker, Michael & Maria Gouskova (to appear). Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic Inquiry* Available as lingbuzz/001622.

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

Becker, Michael, Andrew Nevins & Jonathan Levine (2012). Asymmetries in generalizing alternations to and from initial syllables. *Language* **88**. 231–268.

Beijering, Karin, Charlotte Gooskens & Wilbert Heeringa (2008). Predicting intelligibility and perceived linguistic distance by means of the levenshtein algorithm. In *Linguistics in the Netherlands 2008*, John Benjamins.

Berent, Iris, Colin Wilson, Gary Marcus & Doug Bemis (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry* **43**. 97–119.

Berger, Adam L, Vincent J Della Pietra & Stephen A Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational linguistics* **22**. 39–71.

Berko, Jean (1958). The child's learning of English morphology. *Word* **14**. 150–177.

Blenkiron, Amber & John Alderete (2015). Reduplication in rotuman: A curious case of emergent unmarkedness. In Yuchau Hsiao & Lian-Hee Wee (eds.) *Capturing phonological shades within and across languages*. Newcastle: Cambridge Scholars Publishing, 266–290.

Bobaljik, Jonathan (2000). The ins and outs of contextual allomorphy. In K. K. Grohmann & Caro Struijke (eds.) *University of Maryland working papers in linguistics*, College Park, MD: University of Maryland, vol. 10. 35–71.

Bromberger, Sylvain & Morris Halle (1988). Conceptual issues in morphology.

Broselow, Ellen & John J. McCarthy (1983). A theory of internal reduplication. *The Linguistic Review* **3**. 25–98.

Bybee, Joan (2001). *Phonology and language use*. Cambridge University Press.

Bybee, Joan & Carol Lynn Moder (1983). Morphological classes as natural categories. *Language* . 251–270.

Bybee, Joan & Dan Slobin (1982). Rules and schemas in the development and use of the english past tense. *Language* **58**. 265–289.

Bye, Patrik (2007). Allomorphy: Selection, not optimization. In Sylvia Blaho, Patrik Bye & Martin Krämer (eds.) *Freedom of Analysis?*, Studies in Generative Grammar, Berlin: Mouton de Gruyter.

Churchward, C. Maxwell (1940). *Rotuman grammar and dictionary*. Sydney [Reprinted by AMS Press, 1978]: Australasia Medical Publishing Co.

Dawdy-Hesterberg, Lisa (2014). *The structural and statistical basis of morphological generalization in Arabic*. Ph.D. dissertation, Northwestern University.

Della Pietra, Stephen, Vincent Della Pietra & John Lafferty (1997). Inducing features of random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **19**. 380–393.

Durrett, Greg & Johnn DeNero (2013). Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Elmedlaoui, Mohamed (1995). *Aspects des representations phonologiques dans certaines langues chamito-semitiques*. Rabat: Faculté des lettres et des sciences humaines.

Embick, David (2010). *Localism versus globalism in morphology and phonology*. Cambridge, MA: MIT Press.

Flack, Kathryn (2007). Templatic morphology and indexed markedness constraints. *Linguistic Inquiry* **38**.

Fukazawa, Haruka (1999). *Theoretical implications of OCP effects on features in Optimality Theory*. Ph.D. dissertation, University of Maryland, College Park, MD.

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

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

Gouskova, Maria (2007). The reduplicative template in tonkawa. *Phonology* **24**. 367–396.

Gouskova, Maria & Luiza Newlin-Łukowicz (2014). Selectional restrictions as phonotactics over sublexicons. Lingbuzz/002249.

Hale, Kenneth & Abanel Lacayo Blanco (1989). *Diccionario elemental del Ulwa (Sumu meridional)*. Cambridge, MA: Center for Cognitive Science, MIT.

Halle, Morris (2001). Infixation versus onset metathesis in Tagalog, Chamorro, and Toba Batak. In Kenneth Locke Hale & Michael Kenstowicz (eds.) *Ken Hale: A Life in Language*, MIT Press. 153–168.

Hayes, Bruce (1999a). Anticorrespondence in Yidiɲ. Ms. UCLA.

Hayes, Bruce (1999b). Phonological restructuring in Yidiɲ and its theoretical consequences. In Ben Hermans & Marc van Oostendorp (eds.) *The derivational residue in phonology*, Amsterdam: Benjamins. 175–205.

Hayes, Bruce (2014). Comparative phonotactics.

Hayes, Bruce & Zsuzsa Londe (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* **23**. 59–104.

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

Heeringa, Wilbert (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. dissertation, University of Groningen.

Inkelas, Sharon, C. Orhan Orgun & Cheryl Zoll (1996). Exceptions and static phonological patterns: cophonologies vs. prespecification.

Inkelas, Sharon & Cheryl Zoll (2007). Is grammar dependence real? A comparison between cophonological and indexed constraint approaches to morphologically conditioned phonology. *Linguistics* **45**. 133–171.

Itô, Junko & Armin Mester (1995). The core-periphery structure in the lexicon and constraints on re-ranking. In Jill Beckman, Laura Walsh Dickey & Suzanne Urbanczyk (eds.) *Papers in Optimality Theory*, GLSA. 181–210.

Itô, Junko & Armin Mester (1999). The phonological lexicon. In Natsuko Tsujimura (ed.) *The Handbook of Japanese Linguistics*, Blackwell. 62–100.

Jarosz, Gaja (2006). *Rich Lexicons and Restrictive Grammars - Maximum Likelihood Learning in Optimality Theory*. Ph.D. dissertation, Johns Hopkins University.

Kapatsinski, Vsevolod (2013a). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology.

Language **89**. 110–148.

Kapatsinski, Vsevolod (2013b). Morphological schema induction by means of conditional inference trees. In B. Cartoni, D. Bernhard & D. Tribout (eds.) *TACMO Workshop. Theoretical and Computational Morphology: New Trends and Synergies.*

Kawahara, Shigeto, Kohei Nishimura & Hajime Ono (2002). Unveiling the unmarkedness of Sino-Japanese. *Japanese/Korean Linguistics* **12**. 140–151.

Kiparsky, Paul (2006). The amphichronic program vs. evolutionary phonology. *Theoretical Linguistics* **32**. 217–236.

Kleiweg, Peter (2014). Levenshtein demo [web application]. Http://www.let.rug.nl/kleiweg/lev/.

McCarthy, John J. (1983). Consonantal morphology in the chaha verb. In Michael Barlow, Daniel P. Flickinger & Michael T. Wescoat (eds.) *Proceedings of the West Coast Conference on Formal Linguistics.* 176–188.

McCarthy, John J. (2000). The prosody of phrase in rotuman. *Natural Language and Linguistic Theory* **18**. 147–197.

McCarthy, John J. (2003). OT constraints are categorical. *Phonology* **20**. 75–138.

McCarthy, John J. & Alan Prince (1990). Foot and word in prosodic morphology: The arabic broken plural. *Natural Language and Linguistic Theory* **8**. 209–282.

McCarthy, John J. & Alan Prince (1993a). Generalized alignment. In Geert Booij & Jaap van Marle (eds.) *Yearbook of Morphology*, Dordrecht: Kluwer. 79–153.

McCarthy, John J. & Alan Prince (1993b). *Prosodic morphology I: Constraint interaction and satisfaction.* New Brunswick: Rutgers University Center for Cognitive Science. Available as ROA-482 on the Rutgers Optimality Archive, http://roa.rutgers.edu.

McCarthy, John J. & Alan Prince (1995). Faithfulness and reduplicative identity. In Jill N. Beckman, Laura Walsh & Suzanne Urbanczyk (eds.) *Papers in Optimality Theory*, University of Massachusetts Occasional Papers 18, University of Massachusetts Amherst: UMOP. 249–384.

McCarthy, John J. & Alan Prince (1998). Prosodic morphology. In Andrew Spencer & Arnold Zwicky (eds.) *The Handbook of Morphology*, Basil Blackwell. 283–305.

McCarthy, John J. & Matthew Wolf (2010). Less than zero: Correspondence and the null output. In Curt Rice (ed.) *Modeling Ungrammaticality in Optimality Theory*, London: Equinox.

Moore-Cantwell, Claire & Robert Staubs (2014). Modeling morphological subgeneralizations. In John Kingston, Claire Moore-Cantwell, Joe Pater & Robert Staubs (eds.) *Proceedings of Phonology 2013*, Linguistic Society of America.

Nerbonne, John & Wilbert Heeringa (2010). Measuring dialect differences. In Peter Auer & Jürgen Erich Schmidt (eds.) *Language and Space. An international Handbook of Linguistic Variation*, Mouton de Gruyter, vol. 1. 550–567.

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten & Willem van de Vis (1996). Phonetic distance between dutch dialects. In G. Durieux, W. Daelemans & S. Gillis (eds.) *CLIN VI: Proceedings of the Sixth*

CLIN Meeting, Centre for Dutch Language and Speech (UIA). 185–202.

Nevins, Andrew (2011). Phonologically-conditioned allomorph selection. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.) *The Blackwell Companion to Phonology*, Blackwell.

Paster, Mary (2006). *Phonological Conditions on Affixation.* Ph.D. dissertation, University of California, Berkeley.

Pater, Joe (2000). Nonuniformity in English secondary stress: The role of ranked and lexically specific constraints. *Phonology* **17**. 237–274.

Pater, Joe (2006). The locus of exceptionality: Morpheme-specific phonology as constraint indexation. In Leah Bateman, Michael O'Keefe, Ehren Reilly & Adam Werle (eds.) *Papers in Optimality Theory III*, Amherst, MA: GLSA. 259–296.

Pater, Joe (2008). Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In Steve Parker (ed.) *Phonological Argumentation: Essays on Evidence and Motivation*, London: Equinox. 123–154.

Pater, Joe, Robert Staubs, Karen Jesney & Brian Smith (2012). Learning probabilities over underlying representations. In *Proceedings of the Twelfth Meeting of ACL-SIGMORPHON: Computational Research in Phonetics, Phonology, and Morphology*. Learning probabilities over underlying representations.

Prasada, Sandeep & Steven Pinker (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* . 1–56.

Prasse, Karl (1972). *Manuel de grammaire touarègue (Tahaggart), I–III: Phonétique, écriture, pronom.*. L'Insitut de Copenhague.

Prince, Alan & Paul Smolensky (1993/2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell. [ROA-537].

Rasin, Ezer & Roni Katzir (2013). On evaluation metrics in Optimality Theory. Ms. MIT and TAU.

Riggle, Jason (2006). Using entropy to learn OT grammars from surface forms alone. In Donald Baumer, David Montero & Michael Scanlon (eds.) *Proceedings of WCCFL 25*, Cascadilla Proceedings Project. 346–353.

Seidenberg, Mark S. & David C. Plaut (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science* **38**. 1190–1228.

Smolensky, Paul & Géraldine Legendre (2006). *The Harmonic Mind: From Neural Computation To Optimality-Theoretic Grammar*. MIT Press.

Spruit, Marco René, Wilbert Heeringa & John Nerbonne (2007). Associations among linguistic levels. *Lingua* **119**. 1624–1642.

Tesar, Bruce, John Alderete, Graham Horwood, Nazarré Merchant, Koichi Nishitani & Alan Prince (2003). Surgery in language learning. In G. Garding & M. Tsujimura (eds.) *Proceedings of WCCFL 22*, Somerville, MA: Cascadilla Press. 477–490.

Tesar, Bruce & Paul Smolensky (1998). Learnability in Optimality Theory. *Linguistic Inquiry* **29**. 229–268.

Tesar, Bruce & Paul Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.

White, James (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition* **130**. 96–115.

White, James Clifford (2013). *Bias in phonological learning: Evidence from saltation.* Ph.D. dissertation, University of California, Los Angeles.

Wilson, Colin (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* **30**. 945–982.

Wolf, Matthew (2007). For an autosegmental theory of mutation. In Leah Bateman, Michael O'Keefe, Ehren Reilly & Adam Werle (eds.) *UMOP 32: Papers in Optimality Theory III*, Amherst, MA: GLSA. 315–404.

Zuraw, Kie (2000). *Patterned Exceptions in Phonology*. Ph.D. dissertation, UCLA.

Zuraw, Kie (2007). The role of phonetic knowledge in phonological patterning corpus and survey evidence from tagalog infixation. *Language* **83**. 277–316.