Balinese stem phonotactics and the subregularity hypothesis

Bruce Hayes   Jinyoung Jo

Department of Linguistics, UCLA

November 2020

To appear in *UCLA Working Papers in Linguistics 20*, *“Papers in Phonology”*

Balinese stem phonotactics and the subregularity hypothesis*

ABANDONED

While we enjoyed pursuing this project, we have no further plans to pursue journal publication, since, as an astute reviewer for *Linguistic Inquiry* demonstrated, the point we had hoped to make is mathematically incorrect. The patterning of Balinese stem phonotactics, even though it includes copying, is in fact regular, and this regularity can be captured under a description that a linguist would regard as properly capturing the generalization at hand.

These conclusions arise from a paper by Cohen-Sygal and Wintner (2006) pointed out to us by the *LI* reviewer. Cohen-Sygal and Wintner show that a finite state automaton of the classical type can be associated with a Finite State Registered Automaton (FSRA), which incorporates a set of "registers" forming a kind of memory. The FSRA (specifically, in the variant Cohen-Sygal and Wintner call FSRA*) is capable of expressing bounded copying in a direct and insightful form, thus satisfying the Criterion of Translational Practice in (16) below. Moreover, every FSRA* can be paired with a classical finite state automaton (typically, a very large one), which means that the class of languages recognized by FSRA*s does not go beyond regular — contrary to what we claim in our paper.[2]

Our intent had been to use Balinese as a vivid example in support of two more general points concerning mathematical phonology — points we think are valid nonetheless. First, we endorsed McCollum et al.'s view (2020) that "naive estimates of frequency counts or literature reviews are [not] a reasonable evidential basis for concluding that a particular type of phenomenon is categorically impossible;" see §7.1 below. Second, we suggested (§7.2) that mathematical phonologists should feel empowered to deviate from the straight and narrow paths defined in the textbooks, developing formal deductive systems that are well-adapted to phonological theorizing and description. Indeed, Cohen-Sygal and Wintner's work strikes us as an outstanding example.

*We would like to thank Eric Baković, Jeffrey Heinz, Tim Hunter, Edward Keenan, Adam McCollum, Kie Zuraw, three anonymous reviewers for *Linguistic* Inquiry, and the members of the UCLA Phonology Seminar for helpful comments on earlier drafts of this work.

[2] Roark and Sproat (2007:53-54), whom we cite below, dispute Cohen-Sygal and Wintner's achievement, claiming that finite-state devices are capable of dealing with bounded copying, but only "inelegantly and non-compactly". To the contrary, we are persuaded by the *LI* reviewer's arguments that the work achieves its purpose.

Abstract

In Balinese (Austronesian, Bali), medial consonant clusters generally take the form *nasal* + *obstruent*. This requirement goes largely unenforced, however, in a special class of stems that consist of a repeated, nonmeaningful syllable, e.g. *dapdap* 'kind of tree'. These *pseudoreduplicated* stems contain clusters, such as [pd], that would be aberrant in a normal stem. We analyze the cluster phonotactics of Balinese in detail, demonstrating that the phonology must be able to recognize when a stem consists of copied material. On this basis, we argue that the "subregularity hypothesis," a widely-adopted hypothesis concerning the computational power of phonological systems, is false, and discuss the implications of this finding.

Keywords: Balinese, phonotactics, computational phonology, mathematical linguistics, regular language

Balinese stem phonotactics and the subregularity hypothesis

1. Introduction

A long research tradition (e.g. Johnson 1972, Kaplan and Kay 1994, Frank and Satta 1998) addresses the question of where phonology lies on the hierarchy of grammar complexity, both in its classical Chomskyan version (Chomsky 1956, 1959) and in more recent refinements that distinguish degrees of generative capacity falling below the regular languages (e.g. Rogers and Pullum (2011), Heinz (2011a,b), Heinz and Idsardi (2013)). From this work, a prominent hypothesis has arisen, which we will call the *subregularity hypothesis*, stated for instance in Heinz (2011a:147); it asserts that the computations of the phonological component, in both phonotactics and alternation, fall within the subregular region. In this article, we will suggest that to the contrary, phonology is not even regular; i.e. that the regularity hypothesis for phonology is false, just as it is for morphology and syntax.[3] From this it follows that the subregularity hypothesis is false as well.

The basis of our argument is as follows. We first carry out an analysis of the phonotactics of Balinese (Austronesian, Bali), showing that key phonotactic principles make reference to copying. Next, we examine whether finite state machines (used as a standard criterion for identifying regular languages) can in any meaningful sense generate sets of copied strings, concluding that they cannot. From this it follows that phonology is not regular, and we conclude by discussing the implications of this finding.

[3] The literature for the latter domains is voluminous; good textbook coverage may be found in Roark and Sproat (2007).

2. Background: the phenomenon

The stems of Balinese display an interesting phonotactic pattern noticed by Robert Blust and pointed out in his compendium volume on the Austronesian language family (2009:204). These stems are most often disyllabic, taking either the form CVCVC, or CVCCVC with a medial cluster. Among the latter type, there are two ways to realize the medial CC sequence. Typically, this cluster consists of a nasal homorganic with a following obstruent; i.e. [mp], [nd], [ŋk], etc. However, there also exists a substantial set of stems displaying what Blust calls "fossilized reduplication": the first CVC is identical to the second, as in *bitbit* 'open something a little bit'. The requirement that the medial CC sequence be a homorganic N+C cluster is not enforced in these stems, which instead have rather free patterns of combination for the medial CC. Blust calls these stems "fossilized" because their parts are most often meaningless; thus for *bitbit*, the copied substring *bit* has no meaning and does not exist as a free form in the language — it would not qualify as a morpheme by ordinary criteria. Following Zuraw (2002), who studied a similar case in Tagalog, we will use the term *pseudoreduplicated* for such stems. The following Balinese stems, taken from Barber's dictionary (1979), illustrate the pattern.

(1)*a. "Normal" stems with medial NC*

[mp]	*dampiŋ*	'side, edge'
[nt]	*lontar*	'palm-leaf book'
[ŋg]	*puŋgal*	'to break off'

b. Pseudoreduplicated stems

[pd]	*dapdap*	'tree species'
[gb]	*bugbug*	'pile up'
[ml]	*lumlum*	'yellowish-white'

The pattern "strict NC in normal stems, freedom for pseudoreduplicated stems" is not confined to Balinese; indeed, Blust traces it historically to Proto-Austronesian, the distant ancestor of Balinese spoken thousands of years ago. He also cites other Austronesian languages where this ancient system still survives.

The possible significance of this pattern for applications of formal language theory to phonology is based on the fact that string sets defined by copying cannot be characterized, in the general case, as regular. To establish a closer connection, we must do two things. On the empirical side, it is necessary to give a more detailed analysis of Balinese cluster phonotactics; unsurprisingly, the full pattern is not as clean as the bare description above might imply (see §3); but closer analysis with statistical testing demonstrates that the phonotactics of clusters in pseudoreduplicated stems is indeed distinct from the phonotactics of ordinary stems (§4). Moreover, Zuraw's (2002) earlier study of pseudoreduplication, which we summarize (§5), offers strong cross-linguistic support for our Balinese-specific findings. On the theoretical side, we address in §6 the strictly mathematical issue of whether a bounded string-copying system could be counted as regular, in light of the possibility, put forth by Chandlee (2017) and others, of simply listing every possible copied string. In the remaining section, we suggest some general implications to be drawn from our findings.

3. The Balinese medial clusters

We studied the Balinese clusters by examining a set of stems taken from the massive dictionary (809 pages) compiled from earlier sources and augmented by C. Clyde Barber (1979). We estimate the total number of stems in the dictionary at 29,900, of which about 8100, or 27%, have a medial consonant cluster. Of the latter, about 11% are pseudoreduplicated and 89% are "normal" stems.

In examining this large print corpus, we resorted to random sampling. We collected the "normal" medial-cluster stems on every tenth page of the dictionary, and the pseudoreduplicated medial-cluster stems on every second page; hence we sampled the pseudoreduplicated stems with higher density. This data sample may be examined in the Supplemental Materials for this article.[4]

By definition, the syllables of pseudoreduplicated stems do not occur separately in isolation; i.e. *bitbit* exists, but **bit* does not. This is unsurprising, because only a small minority of stems in Balinese (Beratha (1992:51) estimates 3%) are monosyllabic. There are a modest number of cases in which a reduplicated stem really is morphologically derived: the rare monosyllabic stems sometimes appear in disyllabic reduplicated forms, with some sort of derivational meaning, as in *bək* 'be full', *bəkbək* 'stuffed full'. However, the great bulk of pseudoreduplicated stems (about 89% in our counts) are morphologically underived; i.e. monomorphemic. We excluded the morphologically derived cases from the analysis below.

In (2) we give the phoneme inventory of Balinese, following the analyses of Ward (1973) and Beratha (1992). Symbols have their standard IPA values.

(2) *Balinese phonemes*

a. Consonants

p	t	c	k	
b	d	|	g	
	s			h
m	n	)	N	
	l			

[4] These materials may be obtained from https://linguistics.ucla.edu/people/hayes/.

r

w j

b. Vowels

i u

e ↔ o

a

Table (3) gives counts for each type of medial cluster in the "normal" stems. Here, the first consonant may be read off the row headers and the second off the column headers. The use of boldface and italic type is intended to facilitate reference in the discussion below.

(3) *Counts for medial clusters in "normal" stems*

Second consonant

First consonant	**p**	**t**	**c**	**k**	**b**	**d**	**ɟ**	**g**	**s**	**h**	**m**	**n**	**ɲ**	**ŋ**	**l**	**r**	**w**	**j**
p															*10*	*1*	*1*	
t	1			1							1	1				*7*	*3*	*4*
c															*2*			
k		5							5						*2*	*3*	*1*	
b															*2*	*4*		*1*
d																*3*		*2*
ɟ																		
g							1					1			*2*	*5*		*3*
s	1	16	1	1							1	1				*1*	*3*	
h	1										1							*1*
m	**56**				**73**										1			1
n		**60**		1		**84**												
ɲ			**17**				**24**											
ŋ		1		**83**	1			**50**	**41**	1					5		1	
l					1						1							1
r		5	2		2	1	1	3	4		4	1					1	
w																		
j																		

Let us examine in qualitative terms the generalizations evident in Table (3). We begin by sharpening our earlier description of the *nasal + obstruent* sequences that dominate the set of clusters in "normal" stems. Among these clusters (counts shown in boldface in (3)), the place of articulation of the nasal is predictable as follows. When the obstruent is a stop, as in the examples of (4a) below, the nasal is homorganic to it, just as reported in the simplified description given in §2. When the obstruent of the cluster is the fricative [s], as in (4b), the nasal is still predictable in place, but surprisingly, this is dorsal [N] rather than the expected [n]. The cluster we might actually expect, homorganic [ns], is completely missing from the data.

(4) *Nasal + obstruent clusters in "normal" stems, with predictable place*

a. C_2 *is stop*

[mb]	*sembar*	'spit out of the mouth'
[nt]	*hinten*	'diamond'
[ɲɟ]	*taɲɟal*	'be mischievous'
[ŋk]	*tuŋkak*	'be incomplete'

b. C_2 *is* [s]

[ŋs]	*daŋsək*	'be near'
	taŋsul	'rope, cord'

We can consider this pattern in typological terms. It has been found that nasals show a lesser tendency to participate in place assimilation before fricatives than before stops (Rosenthall 1989, Padgett 1994). Moreover, [N] frequently serves as a default place of articulation for coda nasals (Rice 1996). Thus, typology suggests a sensible account of predictable place for preconsonantal nasals in Balinese: they are homorganic in the context that favors it (before stops) and otherwise take on dorsal place as default.

We turn next to the minority set of clusters in "normal" stems that do not obey Blust's principle. In the upper right quadrant of (3), with counts printed in italic, are various clusters of the form *obstruent + resonant*, where by “resonant” we mean liquid or glide. Examples are given in the left column of (5) below. Such clusters also occur in word-initial position, as shown in the right column.

(5) *Obstruent-resonant clusters*

Cluster	*Medial*		*Initial*	
[pl]	*kuplak*	‘fade, lose color’	*plaŋkan*	‘wooden couch’
[bl]	*geblag*	‘smack with the flat hand’	*blətəŋ*	‘a plant like maize’
[tr]	*setra*	‘tomb, grave’	*trekol*	‘small gun, rifle’
[gr]	*sagrəp*	‘snatch up, seize’	*gremeŋ*	‘in tatters’
[sr]	*hasrama*	‘boarding-house’	*sreŋgen*	‘be angry’
[bj]	*tabja*	‘chili’	*bjasa*	‘ordinary’
[tw]	*satwa*	‘holy’	*twara*	‘poverty’

The fact that these clusters may appear initially suggests that we should treat them as branching onsets; e.g. *kuplak* ‘fade, lose color’ is [ku.plak]. The branching onset analysis is also supported by a modest number of triple clusters, which resolve into a possible coda plus a possible word initial sequence, as in the “normal” stem *ɟumprit* ‘stand on one’s head’, assumed to be [ɟum.prit]; and likewise with a few pseudoreduplicated stems like *blit.blit* ‘bamboo fence’. The key point for present purposes is that the obstruents of obstruent-resonant clusters will not be subject to the constraints on coda consonants to be developed in (10)-(11) below.

The Blustian nasal-obstruent clusters (e.g. (4)) and (secondarily) the medial-onset clusters of (5) form the great bulk (87.5%) of our set of medial clusters in “normal” stems. Aside from these, there are a modest number of clusters distinct from these two types, usually with falling sonority, as well as a few clusters with obstruent sequences. Some examples are given in (6).

(6) *Examples of unusual medial clusters*

[rt]	*murti*	‘excellent, beautiful’
[rm]	*darma*	‘patient, pious’
[rs]	*kursi*	‘chair’
[st]	*nista*	‘be despised’
[ks]	*supeksa*	‘oral declaration in court’
[kt]	*bakta*	‘carry, bring’

Typically, these less-common types occur in the learned lexical strata of Balinese, analogous to the Latinate stratum of English or the Sino-Japanese stratum of Japanese. Barber’s dictionary usually specifies such class membership: the words in question derive from Sanskrit or Kawi, or else are reserved for court or literary usage. We have experimented with modeling the data with the learned words excluded, but since the data become only somewhat more orderly under this procedure, we report only our analysis of the full data set.

Pseudoreduplicated stems, in contrast to “normal” stems, are strikingly free in their medial cluster combinations. While there are systematic gaps in the data (to be discussed in §4 below), the basic generalizations evident in “normal” stems are often violated in pseudoreduplicated stems. Notably (as we will show more carefully below) the *nasal + obstruent* sequences that predominate in “normal” stems seem to be not particularly favored in the pseudoreduplicated stems. This can be seen in coarse-grained terms by comparing Table (3) above for “normal” stems with Table (7) below, which covers pseudoreduplicated stems.

(7) *Counts for medial clusters in pseudoreduplicated stems*

Second consonant

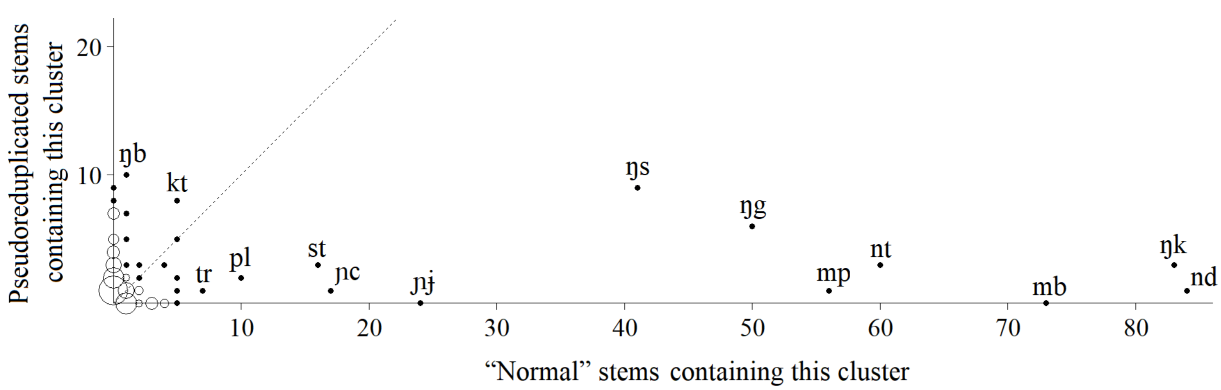
First consonant

	p	**t**	**c**	**k**	**b**	**d**	**ɟ**	**g**	**s**	**h**	**m**	**n**	**ɲ**	**ŋ**	**l**	**r**	**w**	**j**
p		1	3			4	1	1	2				1		*2*	*1*		
t	3		2	1	5	5		1	4				1		*1*	*1*		
c																		
k	7	8	8		7	4	1		5						*2*			
b		1	3	1		1			2						*3*			
d	1	2	2	1	1			1	2						*2*			
ɟ																		
g	2	9	5		3	7	5		7						*1*	*2*		
s	7	3	1	1	4	7		4								*1*		
h	1	1	2	1	2	1	2	1	1				1		*1*			
m	**1**	3	1			1		2	3						1			
n		**3**		2	3	**1**					1				1		1	
ɲ			**1**															
ŋ	2	2	5	**3**	10	3	1	**6**	**9**				1		1			
l	3	2		1	1	4	1	2	2		1		1					
r			1	3	1	1			3									
w																		
j																		

One further fact to observe about the pseudoreduplicated stems is that the identity between the two halves is occasionally incomplete. In 13/288 cases, they have distinct vowels, as in *tigtag* ‘have heated discussion’; and in 81 cases, there is an infix-like element coming after the first C, as in *kəladkad* ‘bamboo tray’, *bletbet* ‘tie’, or *crigcig* ‘walk alone in haste’. We discuss these cases below in §4.2.

We now ask the key question: do these two types of stem, one of them defined by copying, actually have different medial-cluster phonotactics? The question can be addressed intuitively with a graphic display, set up as follows. We create a scattergram in which the dots represent individual clusters like [ŋg], [pl], and so on. We plot each cluster on the scattergram at a location such that the horizontal axis represents the number of “normal” stems that contain this cluster, and the vertical axis represents the number of pseudoreduplicated stems that contain this cluster. If the two strata had the same phonotactic system, then we would expect the frequencies of clusters roughly to match, and we would observe a scatter of points following a diagonal line, though spread out due to random variation. What is actually observed is given in scattergram (8) below. Here, dots that would overlap (cluster sets with identical frequencies for both stem types) are shown by circles whose size reflects the count of overlapping dots, and the $y = x$ diagonal is shown as a dotted line.

(8) *Scattergram: frequencies of cluster types in “normal” and pseudoreduplicated stems*



It can be seen that the scattergram is grossly asymmetrical. Among "normal" stems, the bulk of the data is taken up by stems with Blustian NC clusters, and the remaining clusters are mostly rare. In contrast, among pseudoreduplicated stems there is no obvious preference for Blustian NC clusters; these stems instead distribute their frequency among a great variety of clusters (mostly too dense to label here), so that no one cluster is particularly frequent. This substantial difference will be confirmed quantitatively in the following section, where we turn to formal phonotactic analysis.

To preview where we are headed: non-enforcement of the Blustian cluster principles in the pseudoreduplicated stems implies that the phonological system must somehow "know" which clusters are reduplicated — and hence must be able to detect copying. Before making this claim, however, we will first establish the empirical case more carefully.

4. Analysis of the medial clusters

4.1 Framework

We follow here the MaxEnt (maximum entropy) theory of phonotactics proposed in Hayes and Wilson (2008). This approach employs the MaxEnt version (Smolensky 1986, Goldwater and Johnson 2003) of Harmonic Grammar (Legendre et al. 1990, Legendre et al. 2006), which is itself closely related to Optimality Theory (Prince and Smolensky 1993/2004). In Hayes and Wilson (2008)'s proposal, a phonotactic grammar is assumed to consist of a set of Markedness constraints, each embodying some hypothesized principle in the theory of phonology. Each of these constraints is assigned a *weight*, a real number that expresses its strength; and the weights are computed by fitting the observed frequencies of the members of GEN (often zero) in a data corpus. Using these weights and the pattern of violations, the primary mathematical formula for MaxEnt (reviewed in Hayes and Wilson 2008:383-384) is used to compute for each member of

GEN a probability, which is interpreted as a quantitative characterization of its degree of well-formedness. When the assigned probability is vanishingly small, it implies outright ungrammaticality. Among forms given higher probabilities, differences of probability distinguish nuances of well-formedness, which are typically reflected in corpus frequency. The predictions of the grammar as a whole may be checked by examining how closely it reflects the frequencies of the original corpus, or by running appropriate phonological experiments.

We adopt this MaxEnt approach for two reasons: it renders nuanced distinctions among forms, rather than making a crude up-or-down verdict, and it also permits statistical significance testing of individual phonological constraints. Such testing will permit us to make a more rigorous case that the pseudoreduplicated and "normal" stems of Balinese indeed have distinct phonotactics.

The approach requires a GEN function, for which we adopt here a simple, idealized form, namely a list of the 324 two-consonant clusters that are logically possible given the consonant inventory in (2a).[5] With a simple GEN of this sort, it is not necessary to use custom software to compute the constraint weights and probabilities; indeed a spreadsheet suffices; the utility "Solver" that accompanies Microsoft Excel includes this capacity. Our working spreadsheets, which transparently display our calculations, may be obtained from the Supplemental Materials.

4.2 Lexical strata and REDUP

In the phonological analysis, some means is needed to distinguish pseudoreduplicated from "normal" stems. We suggest that in Balinese we are dealing with *vocabulary strata*, as studied

[5] This means we must discard the 33 triple clusters in our data, for simplicity's sake. These clusters almost always consist of a legal coda followed by a legal branching onset, so accommodating them in a larger-scale model would not be difficult.

e.g. for English by Chomsky and Halle (1968) and for Japanese by Itô and Mester (1995). The latter authors, working, like us, in a constraint-based framework, suggest that phonotactics involves both highly general constraints that hold across the language, as well as others that are stratum-specific.[6] For Balinese we posit Core and Reduplicated strata; and to implement them, we double the list comprising our GEN: 324 candidates for each of the Core and Reduplicated strata.

The Reduplicated stratum of Balinese is defined, of course, primarily by obedience to appropriate principles of copying, for which we follow the theory of pseudoreduplication proposed in Zuraw (2002).[7] Zuraw adapts the widely used Correspondence theory proposed by McCarthy and Prince (1995). She proposes a constraint REDUP, which favors candidates that are parsed into two domains over which a correspondence relation is defined, as in, for instance, [*bit*]$_{\alpha}$[*bit*]$_{\alpha}$. She also adopts a set of correspondence constraints of the "κκ" (copy-to-copy) family, which penalize particular aspects of imperfect copying. Thus, MAX-κκ or DEP-κκ penalizes the [r] of [*crig*]$_{\alpha}$[*cig*]$_{\alpha}$ 'walk alone in haste', and IDENT-κκ(low) penalizes the mismatching vowels of [*tig*]$_{\alpha}$[*tag*]$_{\alpha}$ 'have heated discussion'. Since our focus here is on the medial clusters, which virtually always match, for simplicity we will not include the κκ-correspondence constraints in the analysis, treating all stems as if they matched perfectly. This

[6] For Itô and Mester, the strata must be arranged in concentric form, with the more exotic strata permitting strict supersets of the core stratum. We suggest that this pattern is an accident of the Japanese data that Itô and Mester addressed; e.g. for English both the Latinate and Native strata allow strings that the other stratum would not allow (Hayes 2016); the same pattern as English will be seen below for Balinese.

[7] See Zuraw for extensive discussion of alternative approaches, some of which would also suffice here.

means that we specify the highly-weighted REDUP as the defining constraint of the Reduplicated stratum.

4.3 Inviolable Markedness constraints

In addition to REDUP (Reduplicated stratum only) the legal intervocalic clusters of Balinese are the consequence in our analysis of a set of Markedness constraints. We begin with those that (per Itô and Mester's theory) appear to be shared between strata. As far as we can tell, they are never violated in the attested data.

First, Balinese bans the glides [j, w] in coda (none occurs before a consonant or word-finally; Ward 1973:§2). The consonants of palatal place of articulation ([c, ǀ, ɲ]) are likewise illegal in coda, both before a consonant and word-finally).[8] Further, there are no geminate consonants internal to any stem. Thus, we posit the three exceptionless constraints given in (9).

(9) *Three inviolable constraints*

*GLIDE IN CODA

*PALATAL IN CODA

*GEMINATE

It is appropriate to set these constraints up as "pan-stratal"; that is, not confined to either the Core or the Reduplicated stratum, since there is no advantage to splitting them up. They trivially

[8] An apparent exception is [ɲ] in coda when before homorganic [c, j]. This is a classical phonotactic syndrome crosslinguistically (some coda nasals legal only homorganically), and we assume that the explanation proposed by Itô (1986) would be applicable. The true ban is on *independent* coda place; the place of homorganic nasals is due to a multiply-linked place node, shared with the following stop.

pass the statistical tests outlined below, and they have precedents in other languages.[9] Concerning the weights that should be assigned to them, we encounter the general principle that in MaxEnt, the best-fit weight of a constraint that both explains data and is exceptionless is infinity; and in our spreadsheet implementations, the constraints of (9) receive weights (around 20) that are high enough to give vanishingly small probabilities to violators. The exact value calculated is arbitrary and depends on the search method used.

4.4 CodaCond

The key constraint for present purposes is stated here as a version of Coda Condition (Itô 1986, Prince and Smolensky 1993/2004); it is intended to prefer the canonical Blustian clusters described above under (4).

(10) Coda Condition

The coda of a non-final syllable must be:

—nasal

—pre-obstruent

—homorganic before a stop

—dorsal before a fricative

This is a fair amount of apparatus for a single constraint, and a more principled analysis might attempt to factor it into more parts; the version in (10) should suffice for present purposes. The key point is to test out the constraint against the data of both strata. We will refer to the

[9] Korean, Persian, and English lack glides in coda (though they do have falling diphthongs), Spanish and Korean avoid coda palatals; and geminate-avoiding languages are ubiquitous.

version of this constraint placed in the Core stratum as CODACOND$_{\text{CORE}}$, and the version placed in the Reduplicated stratum as CODACOND$_{\text{REDUP}}$.

The key result is this: in the full Maxent grammar given in (13) below, the best-fit weight for CODACOND$_{\text{CORE}}$ turns out to be 4.4. This is a substantial weight; in particular, one may calculate using the MaxEnt math that a candidate that violates CODACOND$_{\text{CORE}}$ will receive a probability $e^{4.4}$ = 81 times lower than a comparable candidate that obeys it. In contrast, CODACOND$_{\text{REDUP}}$ receives a best-fit weight of only 0.4, corresponding to a probability reduction of just 1.5. Thus, it appears that clusters violating CODACOND are strongly dispreferred in the Core stratum, but only mildly dispreferred in the Reduplicated stratum. We next confirm this result by filling out the constraint inventory and conducting statistical tests.

4.5 Combing through the data for further constraints

In the hope of increasing the reliability of our analysis, we sought additional constraints; we scanned the charts of (3) and (7) for typologically well-supported phonological constraints that also help explain the patterning of the data. The additional constraints that we added are given in (11).

(11) *Other constraints evident in the data*

Constraint	*Reference*	*Description*
*BRANCHING ONSET	(widely adopted; origin unknown)	Avoid syllable onsets with more than one consonant.
*CODA VOICED OBSTRUENT	Kager (1999:40)	Responsible for Final Devoicing in languages such as German. In Balinese, affects *nonfinal* codas

		only.
AGREE(voice)	Lombardi (1999:272)	Consecutive obstruents must agree in voicing, as in Russian.
SYLLABLE CONTACT LAW	Hooper (1976), Murray and Vennemann (1983)	Penalize coda consonants with lower sonority than the following onset.

Along with CODACOND, these are assessed in the section that follows.

4.6 Statistical testing and complete grammar

In a MaxEnt analysis, it is possible to test proposed constraints statistically against the possibility that their apparent effectiveness is merely the result of random variation in the data. Following the method of Hayes, Wilson, and Shisko (2012), we tested our constraints individually with the Likelihood Ratio Test (Wasserman 2004:164). To do this, we compare two grammars, one with all constraints included, the other with just the target constraint excluded; the weights of both grammars are fitted separately to the data. The test yields the degree to which inclusion of the constraint improves the log likelihood of the data, along with a statistical significance value.

Applying the test to the crucial constraint $\text{CODACOND}_{\text{CORE}}$, we find that including the constraint raises the log likelihood of the data from −4544.5 to −3573.7, a difference of 970.7; this corresponds to an encouraging *p*-value of about 10^{-423}. In contrast, $\text{CODACOND}_{\text{REDUP}}$ raises the log likelihood of the data by only 1.7, yielding a p-value of 0.06, which is a nonsignificant result. This is poor performance, but we are nevertheless uncertain whether $\text{CODACOND}_{\text{REDUP}}$ should therefore be excluded from the grammar. In particular, two pseudoreduplicated stems in the corpus are indicated by Barber as optionally modified in obedience to CodaCond$_{\text{Redup}}$: *simsim* ~ *siŋsim* ‘finger-ring’ and *punpun* ~ *pumpun* ‘gather, provide’.

We test all the proposed stratum-specific constraints in this way, and the results are reported in Table (12). “Δ(LogLk)” abbreviates “change in log likelihood arising from inclusion of the constraint.”

(12) *Weights and statistical testing of individual stratum-specific constraints*

Constraint	*Core Stratum*			*Reduplicated Stratum*		
	Weight	*Δ(LogLk)*	*p*	*Weight*	*Δ(LogLk)*	*p*
CodaCond	4.4	970.7	$< 10^{-423}$	0.4	1.7	0.06
*Branching Onset	3.6	647.8	$< 10^{-283}$	1.8	27.6	$< 10^{-12}$
*Coda Voiced Obstruent	1.7	5.2	0.001	0	—	—
Agree(voice)	∞	14.0	$< 10^{-6}$	0	—	—
Syllable Contact Law	0.5	1.8	0.06	2.5	49.6	$< 10^{-22}$

We interpret the results as follows. For six of the ten proposed constraints (CodaCond$_{\text{Core}}$, *Branching Onset$_{\text{Core}}$, Branching Onset$_{\text{Redup}}$, *Coda Voiced Obstruent$_{\text{Core}}$, Agree(voice)$_{\text{Core}}$, and Syllable Contact Law$_{\text{Redup}}$), the *p*-value indicates clear statistical significance, and we include these six constraints in the final grammar given in (13) below. Further, we infer that *Coda Voiced Obstruent$_{\text{Redup}}$ and Agree(voice)$_{\text{Redup}}$ should not be included in the grammar; they have zero weights, implying they have no useful effect in the description of the data. CodaCond$_{\text{Redup}}$ and Syllable Contact Law$_{\text{Core}}$ both test short of significance by conventional standards; however, due to the forms *siŋsim* and *pumpun* just mentioned, we are reluctant to dismiss CodaCond$_{\text{Redup}}$ from the grammar; and for consistency we will apply the same standard to SYLLABLE CONTACT LAW$_{CORE}$, including it as well. We observe that for three of the constraints, the best-fit grammar lists them in both strata, but with different weights.[10]

The statistical testing leads us to adopt a particular grammar, given in (13), as our final hypothesis. This consists of all the constraints that survived after we culled the ones that failed to pass statistical testing.

(13) *A partial phonotactic grammar for Balinese clusters*

Stratal affiliation	*Constraints*	*Weight in Core*	*Weight in Redup.*
Trans-stratal	*PALATAL IN CODA	∞	
	*GLIDE IN CODA	∞	
	*GEMINATE	∞	
Both	CODACOND	4.4	0.4
	*BRANCHING ONSET	3.6	1.8
	SYLLABLE CONTACT LAW	0.5	2.5
Core only	*CODA VOICED OBSTRUENT	1.7	
	AGREE(voice)	∞	
Reduplicated only	REDUP		∞

[10] Further testing, reported in the Supplemental Materials, indicates a statistically significant improvement for the model that lists these constraints in separate strata, relative to a model that merges them across the grammar as we did for the constraints of (9).

4.7 Precautionary analyses

Grammar (13) was constructed by making particular choices about the constraint set, favoring constraints with good typological motivation in order to relate the Balinese cluster system to existing research and show that it is in no way an "exotic" system. However, our choices were to some degree subjective, and so we wish to understand the degree to which our conclusions about Balinese stratal distinctions depend on them. To assess this issue, we therefore also constructed a maximally expressive grammar, which included not only all the constraints of (13), but also a deliberately exhaustive set of 72 segment-specific (unigram) constraints, one for each combination of consonant, stratum, and position (first or second). For instance, [t] IN INITIAL POSITION-CORE receives a mark whenever a [t] occurs in the first position of a cluster in a nonreduplicated stem. The results obtained from this grammar, which may be inspected in the Supplemental Materials, turned out to be very similar to (12)-(13), except that the weight of the $\text{CODACOND}_{\text{CORE}}$ came out somewhat higher, $\text{CODACOND}_{\text{REDUP}}$ somewhat lower. This very rich model also achieves a good fit to the corpus frequencies; $r = .973$; suggesting we have not omitted any especially important cluster constraints.

Another precaution concerns the fact that Balinese, like related Muna (Coetzee and Pater 2008), may tend to avoid homorganic consonants across vowels, e.g. in CVC. For pseudoreduplicated stems, this would tend to reduce the number of medial homorganic NC clusters, since e.g. medial [mp] implies [pVmpVm], which has two homorganic sequences across vowels ([p … m]). In yet another model (Supplemental Materials) we controlled for this by adding to (13) an antihomorganicity constraint in the Reduplicated stratum. The weight of this constraint emerged as very small and nonsignificant; hence we do not consider this as a worrisome confound.

4.8 Semantically vacuous reduplication?

Another approach to our data would be to acknowledge the copying pattern, but attribute it not to phonology but to *semantically vacuous morphology*. In this approach, [dapdap] might be set up as underlying [RED + dap], where RED is the abstract trigger for reduplication under the approach of McCarthy and Prince (1995). /RED + dap/ would then be spelled out in the morphology as [dapdap]. Since copying is reassigned to morphology, the phonology *per se* may be maintained as subregular.

We are reluctant to make use of this strategy because it seem so vulnerable to misuse: the practice of setting up morphemes unconstrained by considerations of meaning makes it possible to say almost anything about the phonotactics of a language and have it come out true. We give two examples.

First, as our digital search indicates,[11] every stem in English that ends in CCC must have [t] or [s] as its third consonant, as in for example *next* [kst], *precinct* [Nkt], *jinx* [Nks], or *glimpse* [mps]. If we set up /-t/ and /-s/ as semantically vacuous suffixes, we can obtain a very clean phonotactics in which no stem at all ends in more than two consonants. The words just cited are not counterexamples, since under the analysis they have CC stems (underlined): [nɛks + t], [prisINk + t], [dZINk + s], and [glImp + s].

Korean has productive compounding, as in [tΣam + ot] ‘sleep-clothes’ = ‘pajamas’. Extension of the compounding principle to semantically vacuous compounding leads to a hitherto unimagined and major simplification of the language’s underlying phonotactics: all stems are monosyllabic, much as in Chinese, with words like [ha + n)l] ‘sky’, [a + p℘ + tΣi] ‘father’.

[11] We used the stem list at https://linguistics.ucla.edu/people/hayes/EnglishPhonologySearch/.

Our two examples are outrageous from the viewpoint of native intuition. The point we wish to make is that, since adopting semantically vacuous morphology leads so readily to absurdly false conclusions elsewhere, it has little credibility as a means of rescuing the subregularity hypothesis.

4.9 Still other analytic alternatives

The proposal we give in §4.1 - §4.6 accounts for the data and is presented to justify our main point, which is that phonological patterns can include copying. However, our account is not the only proposal that can do this. For instance, Eric Baković has suggested to us an alternative using prosodic domains, in which reduplicated forms are given a particular structure, e.g. $(dap)_p(dap)_p$, where $(\ldots)_p$ is some prosodic domain; and CODACOND holds only within $(\ldots)_p$ domains.[12] In this analysis — and indeed any other analysis we can think of (other than pseudo-morphology, already discussed) — it is still necessary for the phonology to be able to recognize that one string is a copy of another.

We mention lastly an analytical alternative that will *not* work, namely merging the two strata and letting REDUP and CODACOND work together to determine the outcome. This singles out a different system, in which the predominant pattern is to obey both constraints, as in real examples such as [dundun] 'be wakened', [guŋguŋ] 'wild strawberry species', or [siŋsiŋ] 'stain'. While Balinese has a *slight* tendency to match this pattern (attested by the small positive weight assigned in our analysis to $\text{CODACOND}_{\text{REDUP}}$), it is just a tendency.

[12] This predicts, in principle, that if the copies are disyllabic, as in *hiŋkəlhiŋkəl* 'laugh very much', then any medial cluster they contain should obey CODACOND, like [ŋk] in the example just given. Unfortunately, most of the copied-disyllable words turn out to be authentic reduplications (with attested bases), so the hypothesis is hard to test.

4.10 Local conclusion

The purpose of the analysis has been to make our key point as carefully as we could: the phonotactics of medial clusters are indubitably distinct in pseudoreduplicated vs. "normal" stems. As (13) shows, the two stem types differ for a number of constraints, particularly for CODACOND, which embodies our characterization of the Blustian medial clusters; these are strongly preferred in "normal" stems but not in pseudoreduplicated stems. To enforce this difference, an adequate phonotactic analysis of Balinese must have access to the information of whether a stem is pseudoreduplicated or not; which implies that the phonotactic assessment in general must include the capacity to detect copied strings. This capacity, in turn, bears on the subregularity hypothesis, in ways to be discussed in §6.

5. Balinese stem phonotactics and "aggressive reduplication"

Our work has been strongly influenced by Zuraw (2002), an article that addresses the issue of phonological copying in more general terms and with further data. We show that this work can be taken as reinforcing our basic conclusion.

Examining data from Tagalog, which is related to Balinese, Zuraw reports a number of findings that match our Balinese results. Pseudoreduplicated stems in Tagalog are abundant, and as in Balinese they permit medial clusters that would not be legal in "normal" stems. Zuraw also cites other instances of this pattern, including cases from languages outside the Austronesian family. For Tagalog, Zuraw argues that the pseudoreduplicated stems are not morphologically derived, and also that the copying relation seen in them is not only present at the underlying level, but is actively maintained in the dynamic phonology, through the suppression of a process of Vowel Raising when it would reduce the similarity of the two CVC portions of the stem. Further, Zuraw offers representative English data showing that individuals who are learning new words often misparse the input so as to render it as two imperfect copies; a characteristic example is *Abu Dhabi*, mislearned as *Abu Dhabu* with matching [ɑbu] strings. This shows that pressure toward imposing a copy relation between parts of a stem — what Zuraw calls “aggressive reduplication” — is present even in a language like English, where pseudoreduplicated stems do not form a large portion of the vocabulary.

In sum, Zuraw’s evidence suggests that if we were to make a guess about the role of Universal Grammar in phonology, it would seem that we would want not to impose a prohibition on copying, but rather a preference for it.

6. Should systems with bounded copying be considered regular?

We next explore the implications of our results for mathematical/computational phonology. We will assume elementary knowledge on the reader’s part of the Chomsky hierarchy and of finite state machines; some good sources for these topics include Roark and Sproat (2007) and Chandlee (2017). Also, in what follows we adopt the common practice of treating regular languages as those that can be recognized by a finite-state machine (Hopcroft and Ullman 1979:29-34).[13]

In assessing the mathematical consequences of copying patterns, it is common to make a distinction between copying of *bounded* strings (upper length limit) and *unbounded* ones. We address the latter first. It is known that unbounded copying processes fall beyond the regular class (for the proof see Hopcroft and Ullman 1979:136). It appears that pseudoreduplication may

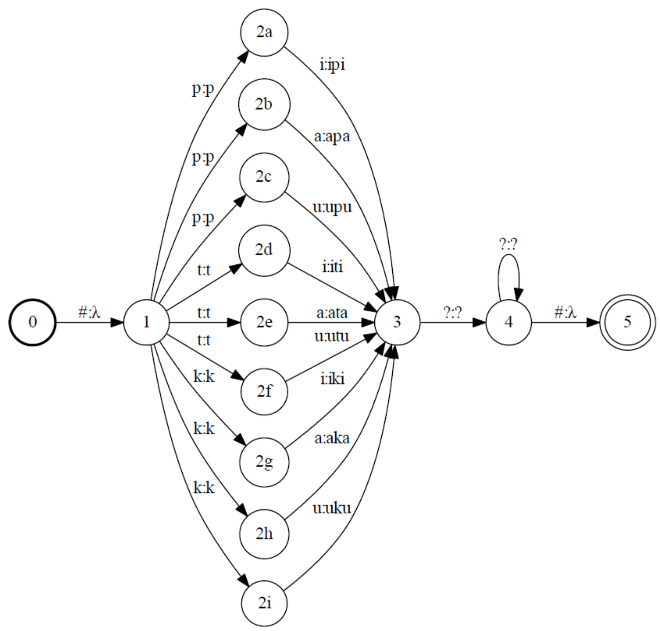
[13] Mohri and Sproat (2006) state that many claims about the computational complexity of languages are *not* valid as theorems, because projecting from a single construction to a whole language is not always a valid inference. They suggest instead that analysts should focus on what sort of automata are capable of recognizing instances of the construction under study, and that is what we are doing here in focusing our attention on finite state acceptors.

indeed be found in unbounded variants. As Zuraw (2002:401) notes, Warlpiri (Pama-Nyungan, Australia; Nash 1980:118–129) appears to be such a case. Beyond this, we suspect that the Zuraw-discovered phenomenon mentioned above of phonological mislearning with erroneous segment-copying extends to polysyllabic forms of English and is almost certainly unbounded. A trisyllabic example we have noticed is *Herdleman and Erdleman*, used for *Haldeman and Ehrlichman* in the United States during the Watergate era.[14]

The question of whether bounded copying — which is what occurs in Balinese — fits in the regular class has been taken up by Roark and Sproat (2007:54-55) and Chandlee (2017:622-623), who suggest that systems of bounded copying may be considered regular provided one provides a suitable characterization of them. For Chandlee, the particular pattern under scrutiny is the common reduplication process that targets the initial CV of a string, as in schematic *pita* ~ *pi-pita*, *kupa* ~ *ku-kupa*. To show that this reduplication is a regular mapping, Chandlee constructs a finite-state transducer that derives the correct outputs in her schematic language. This transducer is set up to include a sufficient number of distinct paths along its arcs to cover *every* possible initial CV sequence, and carries out the copying separately within each path. Supposing, for instance, that the phoneme inventory consists solely of {p, t, k, i, a, u}, then the Chandlean transducer for the *pi-pita* language can be expressed as in (14). The formalism is followed by an informal prose characterization of its behavior.

[14] See books.google.com, search phrase “Herdleman and Erdleman.”

(14) *A finite-state transducer for partial reduplication, after Chandlee* (*2017:623*)



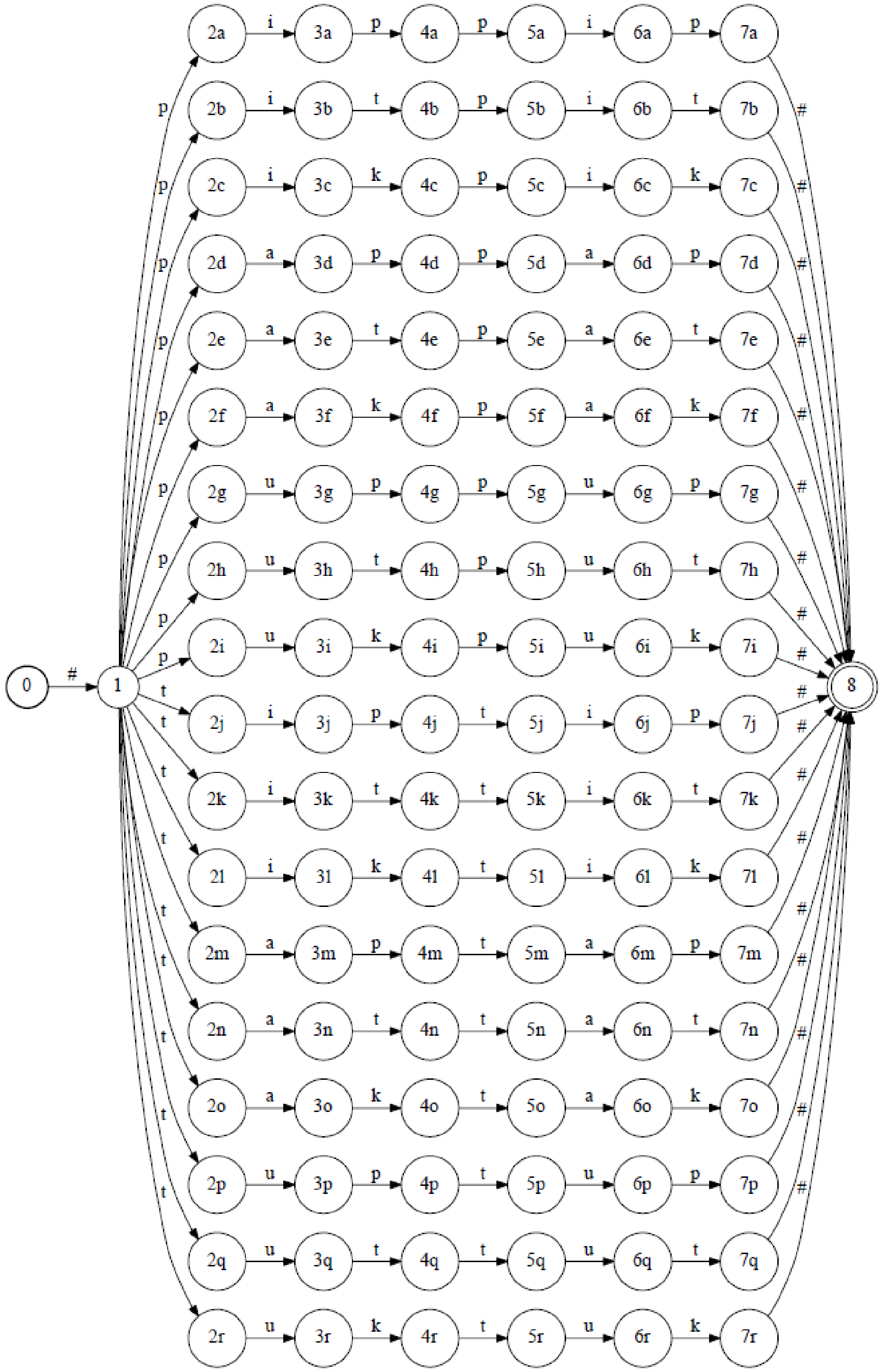
Transitions	*Action*
0 → 1	Navigate the string-start symbol #, transducing it as null (λ).
1 → 2a → 3	If a stem begins [p], followed by [i], first transduce [p] as itself, then replace [i] with [ipi].
1 → 2b → 3	If a stem begins [p], followed by [a], first transduce [p] as itself, then replace [a] with [apa].
1 → 2c → 3	If a stem begins [p], followed by [u], first transduce [p] as itself, then replace [u] with [upu].
etc.	(same, for six more cases)
3 → 4, 4 → 4	Transduce all remaining segments as themselves (multiple arcs, abbreviated as one).

4 → 5	Navigate the string-termination symbol #, transducing it as null.

A similar analysis, covering partial reduplication in Gothic, has been put forth by Roark and Sproat (2007:53-55). We note that the latter authors express considerable distaste for their own account, calling it both “naïve” and “clearly inelegant”. Below, we will give a reason for holding an even stronger negative opinion.

The Chandlee/Roark/Sproat strategy can be applied to Balinese stem structure. Here, we are dealing with phonotactics, so a finite-state acceptor, rather than transducer, is appropriate. In (15) we give an acceptor along the lines (14) that would work for a miniature version of Balinese with the segment inventory {p, t, k, i, a, u}; stems starting with [k] have been omitted for brevity.

(15) *A finite-state acceptor for Balinese pseudoreduplicated stems (6-phoneme inventory)*



As can be seen, the application of the strategy just given for Balinese pseudoreduplicated stems would consist essentially of listing all the logical possibilities individually.[15] For this reason we will call the analytic gambit put forth by Chandlee, Roark, and Sproat the *full-listing* strategy.

6.1 Evaluating the full-listing strategy

We would not deny that the finite state machines employed in the full-listing strategy are valid instances of their formal type. However, the full-listing strategy raises issues — often kept implicit — concerning the methodology employed in the computational analysis of linguistic systems; that is to say, how scholars agree to accept a particular automaton as a formal rendering of a linguistic pattern. Insofar as mathematical linguistics is intended to shed insight on linguistic questions, we think these assumptions are worth articulating.[16]

For the present case, a good role model can be found in computational work on formal syntax. In this area, one assumption seems very firm, if seldom articulated; we give an informal rendition in (16), refining it as we proceed.

---

[15] Acceptor (15) could be made somewhat smaller by collapsing together certain sets of nodes numbered 2, 3, 6, and 7; we have kept the uncollapsed version here for legibility. The point at hand would not be affected.

[16] For views on similar lines see Chomsky (1957:ch. 5), Culy (1985:350), and Savitch (1993). Dassow et al. (1997) write, "[concerning the question of] where the natural languages are placed in the Chomsky hierarchy … the debate started in 1959 and is not settled. Various arguments over English, Mohawk, Swiss German, Bambara, Chinese, etc., were given, refuted, rehabilitated … The main difficulty is not a mathematical one but a linguistic one."

(16) *Criterion of translational practice*

A formal grammar intended to describe a linguistic pattern is expected to express the same pattern under additions to the vocabulary (i.e., to the alphabet of terminals).

Introductory texts commonly rely on (16) when they introduce context-free grammars: the exemplification of the production rules introducing terminal vocabulary is normally rather skimpy, for it is assumed that the reader could easily provide appropriate additional production rules to cover novel vocabulary. Thus, Hopcroft and Ullman (1979:78) give N → *boy* as the only rule introducing nouns, and invite the reader to add others as appropriate.

However, (16) is not just a basis for expository simplification; it reflects a deeper empirical point about language, well understood by linguists: languages often expand their set of syntactic terminals, for instance with loanwords, and the novel words respect the existing syntactic principles of the language. From this, we see that what a grammar expresses is a general *pattern* in a language, and the set of vocabulary items that can be used to embody the pattern is only an incidental fact, changing over time even within a single idiolect.

Taking this point of view, it becomes a matter of interest, for each mathematical class of grammar, in what ways the grammar can be extended with the addition of vocabulary while still preserving its characterization of the linguistic pattern. For context-free grammars, we suggest that the method implicit in current practice is to limit the expansion of a grammar to adding “clones” of the existing production rules that introduce terminal symbols. For instance, if a context-free grammar already contains the production rules N → *spaghetti* and N → *linguine,* then the introduction of a novel word, say *strappatelle,* would be accommodated by cloning one of these rules to create N → *strappatelle*. This keeps the overall syntactic pattern intact, and lets *strappatelle* be distributed according to the existing principles applicable to nouns.

Turning to phonology, we observe that expansion of the vocabulary also occurs here. Languages frequently acquire new phonemes, often through loanwords, and typically the existing phonological pattern is extended to these phonemes following the natural classes to which they belong. Thus, Wiese (1996:200-201) points out that the novel phoneme /ʒ/ of German undergoes Final Devoicing, just like the established voiced obstruents of the language. Halle (1978:301-302, citing Menn) and Pinker (1999:94) draw implications from the behavior of the name of the German composer *Bach*: here, the final segment, faithfully rendered as [x] by some English speakers, triggers the voiceless allomorph of the past tense ([-t], as in *Handel out-Bach*[t] *Bach*) and the non-syllabic voiceless allomorph of the plural, as in *Bach*[s]; these outcomes reflect the status of [x] as a member of the natural class of voiceless non-sibilants. Extension of the segment inventory may also involve copying; thus Zuraw (1996:9) demonstrates extension of CV- reduplication to novel segments in Tagalog, whose speakers extend it to segments like [T] as in *thank you*, inflected in Tagalog as [mag-Tɛ-TæNkju];[17] and Berent et al. (2002) experimentally demonstrate the ability of Hebrew speakers to extend patterns of templatic copying to the non-Hebrew sounds [θ], [tʃ], [dʒ], and [w].

We suggest, therefore, that proposed finite-state implementations of phonological generalizations should be required to respect the same criterion (16) established in syntactic work; i.e. that the system should continue to express the same generalizations under expansions of the vocabulary. In the Appendix to this article, we work out a simple formal approach to this

[17] The difference in vowels is suggested by Zuraw to reflect allophonic variation.

task for both context-free grammars and finite state machines. For the latter, the criterion developed there is given below in (17).

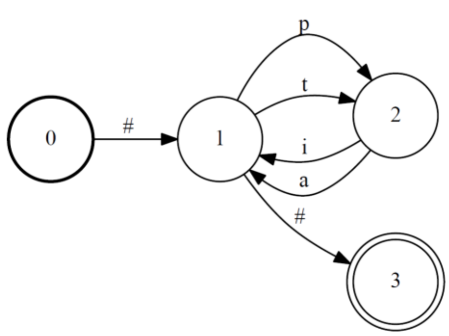
(17) *Criterion of translational practice for finite-state machines*

In expanding a finite-state machine to accommodate novel terminals, the only permitted change should be to add new transitions to already-connected state pairs, in an existing direction.

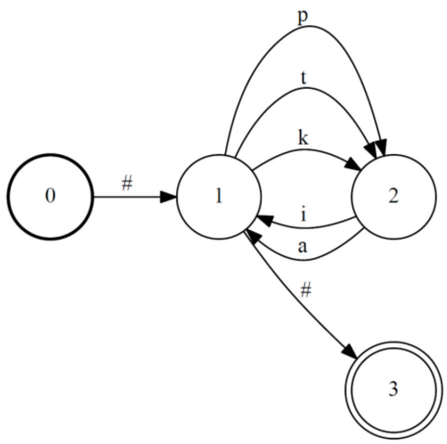
The Appendix shows that this is actually the *same* criterion that is commonly applied to context-free grammars, as just discussed.

We demonstrate (17) with a simple example. Imagine a language with phoneme inventory [p, t, i, a], where every word consists of a sequence of one or more CV sequences, which could be imagined to be syllables; thus [pi], [tipa], [tapiti], etc. A simple finite-state acceptor for this language is given in (18a); it can be adapted to include a hypothetical loan phoneme [k] by adding a new transition arc for [k] in an existing direction (1 → 2), as in (18b).

(18)a. *Finite-state acceptor for a CV language with inventory* [p, t, i, a]

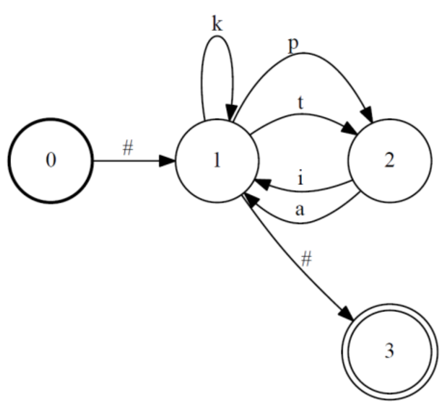


b. *Expansion of this acceptor to include* [k]



This augmentation passes our test, and indeed the new grammar expresses the same phonological generalization, except that [k] is now included in the inventory of consonants. In contrast, if we were we to add [k] in violation of (17) — say, at the location 1 → 1, which has no pre-existing arcs — we would create the bad generalization in (19).

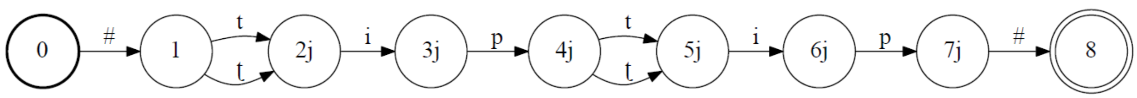
(19) *Impossible generalization of (18a)*



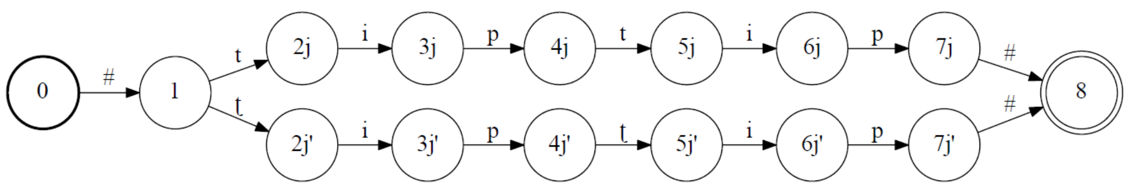
This wrongly introduces an entirely new pattern in the phonology, namely syllables like *[kkpa], beginning with strings of [k].

Assuming criterion (17), we can return to the problem of copying. We assert that the "full-listing" treatment of copying in general, and more specifically the application of it to Balinese phonotactics in (15), runs afoul of criterion (17). This is because, as soon as we agree to limit the treatment of vocabulary expansion to adding novel arcs in existing locations, the grammar would lose the ability to copy. For example, taking just the path labeled "j" in (15), and adding new arcs for retroflex [ʈ] (a hypothetical new phoneme[18]) as in (20a), we would permit not only the legal pseudoreduplicated forms [tiptip] and [ʈipʈip], but also the illegal forms *[ʈiptip] and *[tipʈip]. The only workable expansion would be one that added new states (20b), in violation of criterion (17).

(20) *a. Failed version adding only new arcs; permits* *[ʈiptip], *[tipʈip]



b. Version that works, violating criterion (17)



We sum up our discussion of bounded copying and regularity as follows. We have suggested that acceptable renderings of linguistic patterns in mathematical form should be subject to the criterion of translational practice given in (16), operationalized for finite-state machines as in (17). If one accepts this criterion, the full-listing account of bounded copying is not just "inelegant," per Roark and Sproat, but should be excluded entirely, for *it does not describe the intended language* —it describes copying, but not as a pattern. From this it would follow that even bounded-copying systems like Balinese should not be considered regular.

[18] For Balinese, this may not be so hypothetical; from orthographic evidence (Barber 1979) we know that [ʈ] was once a borrowed phoneme of this language, and it occurred in pseudoreduplicated forms. Subsequent sound change has removed [ʈ] from the Balinese phoneme inventory, merging it with [t].

7. Discussion

We conclude that the data from Balinese and other languages support an analysis in which the phonological grammar must provide for copying, thus constituting an exception to the regularity hypothesis. From this it follows that the subregularity hypothesis is not correct either. We offer below some implications for future research that are related to our finding.

7.1 The "argument from silence" and statistical testing

The subregularity hypothesis bears on an important current methodological dispute. Generative linguists have for decades made use of a research method that might be called the "argument from silence." It works as follows: when the research community is unaware of the existence of any examples of phenomenon X, this is taken as appropriate empirical support for a formal theory that categorically excludes X. McCollum et al. (2020) articulate this assumption as in (21).

(21) *The argument from silence, per McCollum et al. (2020, §6.3)*

> "Naive estimates of frequency counts or literature reviews are a reasonable evidential basis for concluding that a particular type of phenomenon is categorically impossible."

The argument from silence is often expressed not as an observation about the data, but about the *theory* —the theory is said to be "restrictive." Philosophically, it is hard to disagree with the view that restrictiveness offers explanatory force and is *a priori* desirable.

However, McCollum et al. suggest that, for purposes of research, it would wise to reverse the burden of proof: the advocate of the view that a principle of Universal Grammar underlies typology should not just assess formal restrictiveness, but also carry out statistical testing to assess the credibility of the currently available data. In McCollum et al.'s words: "any offered linguistic universal should be presented alongside an explicit measure of evidential strength."

The advance of Bayesian statistical techniques in recent years means that it has become more feasible to evaluate cases of the argument from silence by means of statistical testing. For instance, it can now be estimated just how many languages would be needed to examine to establish with a credibly high probability that the absence of phenomenon X is systematic and not accidental. Piantadosi and Gibson (2014) offer an analysis along these lines, and suggest that the number of languages needed to achieve serious credibility can be quite high, often in the hundreds. To our knowledge, no testing on this scale has taken place for the subregularity hypothesis in phonology.[19]

McCollum et al. further suggest that there is a natural consequence of not providing statistical tests in support of restrictiveness proposals: since the possibility that they are true by accident has not been explicitly dealt with, we are likely sooner or later to be confronted with counterexamples. The main empirical contribution of McCollum et al.'s paper is to bring forth several counterexamples to Jardine's (2016) claim that tonal phonology is computationally richer than segmental phonology (specifically, only tone requires non-deterministic regular functions). Our Balinese example, counterexemplifying the regularity hypothesis, is similar; and we think it also provides support for McCollum et al.'s critique of (21) as a research practice.

7.2 Working with new axioms and definitions

Of course, not all applications of mathematical phonology involve the "argument from silence" and its issues. Notably, scholars have employed formal language theory as the basis of learning algorithms (e.g. Heinz 2010, Jardine and Heinz 2016, Chandlee 2017:601), which are readily

[19] One reason for this lack may be that most of the work of developing appropriate statistical techniques for testing restrictiveness hypotheses remains to be done. It is tempting to suggest that members of the research community in computational linguistics, who have strong qualifications, might take on this task.

tested empirically by examining the generalizations they learn from real-language corpora (Wilson and Gallagher 2018, Jarosz 2019). Further, it is possible to carry out experiments that test whether certain patterns that fall high on the complexity hierarchy are easily learned by experimental subjects; see for instance Lai (2015), McMullin and Hansson (2019), and Avcu and Hestvik (2020).

For present purposes we adopt the premise that such research is likely to find that copying — which is ubiquitous in language[20] — is not a particularly great challenge to human learning or processing. In other words, we conjecture that copying really does represent a clash between putative mathematical complexity and linguistic ordinariness.[21]

We see this situation as a research opportunity, for the following reason. It is an error, we suggest, to suppose that the categories of the grammatical complexity hierarchy, as given in textbooks on formal language theory, have any *a priori* claim to match with the complexity principles that govern human language structure and language learning. Specifically, if standard formal language theory places copying anomalously high in its asserted complexity, then a sensible response would be, invent a different formal language theory, equally rich in theorems and proofs.

[20] Aside from pseudoreduplication, aggressive reduplication, morphological reduplication, and syntactic copying we note that rhyme and alliteration are widespread in the world's versification systems and likewise require copy-detection.

[21] We are not the first to make such a suggestion. Öttl et al. (2015) argue, based on earlier work and their own experiments, that cross-serial syntactic dependencies (mildly context-sensitive) are easier to learn than nested dependencies (context-free); "formal complexity and cognitive complexity are not two sides of the same coin" (p. 14).

Our suggestion is reasonable because formal language theory is an axiomatic system, whose theorems are deduced from a particular initial set of axioms and definitions. In any such system, the initial choices are made by the analyst, and are not eternal verities.[22] In the present case, there is no particular reason to assume that the choices made by the founders of modern formal language theory will necessarily be the ones that turn out to be insightful when applied to phonology.

To be more specific, Hopcroft and Ullman's textbook (1979:28) defines the regular languages as the result of recursive application of specific principles of concatenation and closure. From this, it can ultimately be proven (p. 136) that languages with unbounded copying are not regular. Indeed, languages with copying end up quite high in the standard formal language hierarchy, suggesting great complexity. However, it would also have been possible — plausible, in our view — to include copying as a primitive operation in the original set of definitions, which might have led to a different, and perhaps more empirically applicable, hierarchy of language types.

The idea of starting out the formalized system with different definitions has some history. In *The Sound Pattern of English* (1968), Chomsky and Halle participated in mathematical phonology by including a rigorous formalization (pp. 390-399) of the phonological theory employed in their analysis. Interestingly, the primitives of the Chomsky-Halle system are features, with segments defined as a derivative notion. A positive consequence of this choice is

[22] The text gives the modern view, but this view was not always the dominant one; it was the development of non-Euclidean geometry in the nineteenth century that led people to understand that axioms are choices made by the analyst, rather than eternal truths. A lively narrative of this important intellectual shift is offered in Bell (1937:294-306).

that, from the beginning, their system is responsive to the phenomena of natural classes and pattern symmetry that are so widely observed in phonology — and only seldom addressed in current formal language study.[23] Beyond this, the inventory of axiomatic systems already includes varieties responsive to copying in some degree; and these are perhaps underexplored as theories of linguistic systems.[24]

In sum, we are suggesting that the current narrow focus within mathematical phonology on textbook formal language theory may be counterproductive; and that developing proof-theoretic formal systems from the bottom up that are more closely tailored to phonological phenomena might in the long run be a more fruitful path, finding greater empirical applicability and greater integration with phonological theory.

Appendix: defining normative translational practice for context-free grammars and finite-state machines

In (16) above we presented a normative principle, implicit in existing work, that a formal grammar intended to describe a linguistic pattern must express the same pattern under additions to the vocabulary. To make this principle explicit throughout mathematical linguistics is a major enterprise, for which a serious start has been made by Keenan and Stabler (2003). In the present

[23] Exceptions include Eisner (1997) and Albro (2004).

[24] A notable effort is Dolatian and Heinz (2018), which treats copying with a higher-order type of automaton, the two-way finite state transducer, for which some proof-theoretic work already exists (p. 68). However, their study is not directly applicable to the Balinese case, since two-way finite state transducers can only treat reduplication as a active copying operation (useful for, e.g., morphology), not as an acceptor (needed for phonotactics); see pp. 68, 73.

context, it will suffice to be explicit regarding just two elementary grammar types, context-free grammars and finite-state automata.

A.1 Context-free grammars

Per Hopcroft and Ullmann (1979:79), let G be a context-free grammar (V, T, P, S), where V and T are disjoint sets of nonterminal and terminal elements, P is a set of production rules (in which nonterminals are expanded as sequences of nonterminals and terminals), and S is the start symbol. We define a **lexical extension** of G as in (22):

(22) *Defn.: lexical extension*

Let $A \rightarrow \alpha\ w\ \beta$ be some production rule of G as defined above. Let w' be a symbol not in T or V. Then the context-free grammar $G' = (V, T \cup w', P \cup A \rightarrow \alpha\ w'\ \beta, S)$ is a **lexical extension** of G.

Further, if G′ is a lexical extension of G, then any lexical extension of G′ is also a lexical extension of G (recursive definition).

What this definition says is that if a context free grammar includes the production rule $A \rightarrow \alpha\ w\ \beta$, then for a new terminal symbol w' the production rule $A \rightarrow \alpha\ w'\ \beta$ may be added to it to form a lexical extension, and that this process of adding terminals may be continued ad libitum.[25] In such a process, the *non*-terminals in the rules are retained unaltered, which means

[25] This differs slightly from the approach in the main text, in which only rules of the form $A \rightarrow w$ (A has a single daughter) are cloned. It would work acceptably to define cloning in (22) to create only $A \rightarrow w'$ from $A \rightarrow w$; and this would indeed match more closely with the usual practice of linguists. We maintain the more general definition in (22) since it is needed to cover the extension to finite-state machines given later on.

that the fundamental distributional generalizations, which are governed by the non-terminals, will remain unaltered.

Using the concept of lexical extension we can restate more precisely the principle of normative practice given in (16), namely as (23).

(23) *Criterion of translational practice, restated*

A grammar should not be accepted as a formalization of a linguistic pattern if its lexical extensions fail to manifest the same pattern.

Here are examples: (a) the lexical extensions of context-free grammars in which VP can take a maximum of two NP objects likewise are grammars in which VP can only take a maximum of two objects, since the crucial production rule VP → V NP NP cannot be altered in forming a lexical extension, nor can any other legal change increase the number of possible NP daughters of VP. (b) The lexical extension of the grammar $A \rightarrow aAa$, $A \rightarrow bAb$, $A \rightarrow \epsilon$, which generates palindromes, does not generate palindromes, since adding the production rule $A \rightarrow aAc$ breaks palindrome-matching.[26]

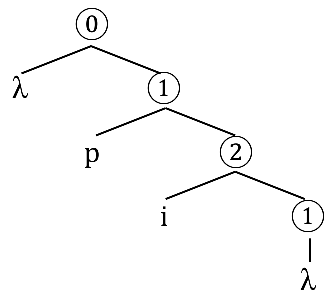
A.2 Finite-state machines

As defined in (22), lexical extension can be extended to finite state machines by using a well-known theorem (Hopcroft and Ullman 1979:217-220) stating that any finite-state machine F can be expressed as a right-linear grammar G. A right-linear grammar is a restricted form of context-free grammar in which all the production rules are of the form $A \rightarrow w\ B$ or $A \rightarrow w$. For instance,

[26] Since natural languages evidently do not deploy palindromic phenomena, we take this in principle to be a *good* result; the criterion of translational practice in this particular case beneficially trims back overgeneration.

the finite-state machine in (18a) can be expressed as right-linear grammar containing six production rules, one for each arc (λ represents null): ⓪ → λ ①, ① → p ②, ① → t ②, ② → i ①, ② → a ①, ① → λ. The production rules derive an output string in a way that recapitulates the path taken through the finite state machine as it emits the same output; for instance, (24) gives the output tree for [pi]:

(24) *Tree for* [*pi*], *generated by a right-linear grammar equivalent to (18a)*



Because of the equivalence theorem, we can take a finite-state machine F, translate it into its context-free counterpart G, use (22) to generate its lexical extension G′, and lastly translate G′ back into the finite-state-machine F′. We will say that if such a string of operations is carried out, then F′ is a lexical extension of F.

Lastly, we operationalize this definition so we can apply it directly to the finite state machines. It should be clear from the above that when we translate a right-linear grammar with production rule A → *w* B into its finite state equivalent, the counterparts of A and B are states, and *w* is the symbol emitted when traversing the arc connecting A and B. From this it follows that the definition in (17) in the main text, which permits novel terminals to be added only as the labels of novel arcs connecting existing states (in the same direction), identifies the lexical extensions of a finite state machine, and thus serves as an adequate basis for normative practice as defined in (23).

A final note: principle (17) represents what we think is a *necessary* condition for validating an automaton as a translation of a linguistic system. However, it is hardly a *sufficient* condition. For instance, when novel terminals are added to a grammar, normative practice is that they be included only in rules that introduce other items of the same natural class. Thus, in expanding (18a) above it would be disastrous to add [k] to the set of arcs for vowels, rather than the set of arcs for consonants, since [k] is itself a consonant. The formalization of such restrictions is important but goes beyond what is needed in this context. The necessary condition we have established already suffices to identify violations of normative practice, in particular the full-listing account of phonological copying.

References

Albro, Daniel. 2005. Studies in computational Optimality Theory, with special reference to the phonological system of Malagasy. Ph.D. dissertation, University of California, Los Angeles, Los Angeles.

Avcu, Enes, and Arild Hestvik. 2020. Unlearnable phonotactics. *Glossa: a journal of general linguistics* 5, no. 1 (2020).

Barber, C. Clyde. 1979. *A Balinese-English dictionary*. Aberdeen: University of Aberdeen.

Bell, E. T. (1937) *Men of Mathematics*. New York: Simon and Schuster.

Beratha, Ni Luh Sutjiati. 1992. Evolution of verbal morphology in Balinese. Ph.D. dissertation, Australian National University, Canberra.

Berent, Iris, Gary F. Marcus, Joseph Shimron, and Adamantios I. Gafos. 2002. The scope of linguistic generalizations: evidence from Hebrew word formation. *Cognition* 83:113–139.

Blust, Robert. 2009. *The Austronesian languages*. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University.

Chandlee, Jane. 2017. Computational locality in morphological maps. *Morphology* 27:599–641.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2:113–124.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1959. On certain formal properties of grammars. *Information and Control* 2:137–167.

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

Coetzee, Andries W., and Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 26:289–337.

Cohen-Sygal, Yael and Shuly Wintner. 2006. Finite-state registered automata for non-concatenative morphology. *Computational Linguistics* 32:49-82.

Culy, Christopher. 1985. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy* 8:345–351.

Dassow, Jürgen, Gheorghe Păun, and Arto Salomaa. 1997. Grammars with controlled derivations. *Handbook of Formal Languages*, ed. Grzegorz Rozenberg and Arto Salomaa. Berlin: Springer.

Dolatian, Hossep, and Jeffrey Heinz. 2018. Learning reduplication with 2-way transducers. *Proceedings of Machine Learning Research* 93:67–80.

Eisner, Jason. 1997. Efficient generation in primitive Optimality Theory. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 313-320.

Frank, Robert, and Giorgio Satta. 1998. Optimality theory and the generative complexity of constraint violability. *Computational Linguistics* 24: 307–315.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 133−122.

Halle, Morris. 1978. Knowledge unlearned and untaught: What speakers know about the sounds of their language. In *Linguistic theory and psychological reality*, ed. by Morris Halle, Joan Bresnan and George Miller, 294–303. Cambridge, MA: MIT Press.

Hayes, Bruce. 2016. Comparative phonotactics. In *Chicago Linguistic Society (CLS) 50*, ed. by Ross Burkholder, Carlos Cisneros, and Emily R. Coppes. 265–285. Chicago: Chicago Linguistic Society.

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

Hayes, Bruce, Colin Wilson and Anne Shisko (2012) Maxent grammars for the metrics of Shakespeare and Milton. *Language* 88:691‒731.

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

Heinz, Jeffrey. 2011a. Computational phonology – Part I. Foundations. *Linguistics and Language Compass* 5:140–152.

Heinz, Jeffrey. 2011b. Computational phonology – Part II: Grammars, learning, and the future. *Language and Linguistics Compass* 5:153–168.

Heinz, Jeffrey, and William Idsardi. 2013. What complexity differences reveal about domains in language. *Topics in Cognitive Science* 5:111–131.

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

Hopcroft, John E., and Jeffrey D. Ullman. 1979. *Introduction to automata theory, languages, and computation*. Reading, MA: Addison-Wesley.

Itô, Junko. 1986. Syllable theory in prosodic phonology, Ph.D. Dissertation, University of Massachusetts Amherst, Amherst.

Itô, Junko, and Armin Mester. 1995. Japanese phonology. In *The handbook of phonological theory,* ed. by John Goldsmith, 817–838. Oxford: Blackwell.

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

Jardine, Adam. 2016. Computationally, tone is different. *Phonology* 33:247–283.

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

Johnson, Douglas. C. 1972. *Formal aspects of phonological description*. The Hague: Mouton.

Kager, René. 1999. *Optimality Theory*. Cambridge: Cambridge University Press.

Kaplan, Ronald M., and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20:331–378.

Keenan, Edward, and Edward Stabler (2003) *Bare grammar: Lectures on linguistic invariants*. Stanford, CA: Center for the Study of Language and Information.

Lai, Regina. 2015. Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry* 46:425–451.

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

Legendre, Géraldine, Antonella Sorace, and Paul Smolensky. 2006. The Optimality Theory–Harmonic Grammar connection. In Paul Smolensky and Géraldine Legendre (eds.), *The Harmonic Mind*, 339–402. Cambridge, MA: MIT Press.

Lombardi, Linda. 1999. Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language and Linguistic Theory* 17:267–302.

McCarthy, John J., and Alan Prince. 1995. Faithfulness and Reduplicative Identity. University of *Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*, ed. by Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey, 249–348. Amherst: University of Massachusetts, Department of Linguistics.

McCollum, Adam., Eric Baković, Anna Mai, and Eric Meinhardt. 2020. Unbounded circumambient patterns in segmental phonology. *Phonology* 37:215-255.

McMullin, Kevin and Hansson, Gunnar (2019) Inductive learning of locality relations in segmental phonology. *Laboratory Phonology* 10:1–53.

Mohri, Mehryar, and Richard Sproat. 2006. On a common fallacy in computational linguistics. *SKY Journal of Linguistics* 19:432–439.

Murray, Robert W., and Theo Vennemann. 1983. Sound change and syllable structure in Germanic phonology. *Language* 59:514–528.

Nash, David. 1980. Topics in Warlpiri grammar. Ph.D. dissertation, MIT, Cambridge, MA.

Öttl, Birgit, Gerhard Jäger, and Barbara Kaup (2015) Does formal complexity reflect cognitive complexity? Investigating aspects of the Chomsky hierarchy in an artificial language learning study. *PLoS ONE* 10(4):e0123059.

Padgett, Jaye. 1994. Stricture and nasal place assimilation. *Natural Language and Linguistic Theory* 12:465–513.

Piantodosi, Steven T. and Edward Gibson. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38:736-756.

Pinker, Steven. 1999. *Words and Rules*. New York: Basic Books.

Prince, Alan, and Paul Smolensky. 2004 [1993]. *Optimality theory: Constraint interaction in generative grammar*. Cambridge, MA: Blackwell.

Rice, Karen. 1996. Default variability: The coronal-velar relationship. *Natural Language and Linguistic Theory* 14:493–543.

Roark, Brian, and Richard Sproat. 2007. *Computational approaches to morphology and syntax*. Oxford: Oxford University Press.

Rogers, James, and Geoffrey Pullum. 2011. Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information* 20:329–342.

Rosenthall, Samuel. 1989. The phonology of nasal-obstruent sequences. M.A. thesis, McGill University.

Savitch, Walter J. 1993. Why it might pay to assume that languages are infinite. *Annals of mathematics and artificial intelligence* 8:17–25.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.*

Volume 1: Foundations ed. by David. E. Rumelhart, James. L. McClelland and the PDP Research Group Cambridge, MA: MIT Press/Bradford Books. 194–281.

Ward, Jack. 1973. Phonology, morphophonemics and the dimensions of variation in spoken Balinese. Ph.D. dissertation, Cornell University, Ithaca.

Wasserman, Larry. 2004. *All of statistics: A concise course in statistical inference.* New York: Springer.

Wiese, Richard. 1996. *The phonology of German*. Oxford: Clarendon Press.

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

Zuraw, Kie. 1996. Floating phonotactics: Infixation and reduplication in Tagalog loanwords. M.A. thesis, University of California, Los Angeles, Los Angeles.

Zuraw, Kie. 2002 Aggressive reduplication. *Phonology* 19:395–439.