Oxford
LINGUISTICS

Rules, Constraints, and Phonological Phenomena

Edited by
Bert Vaux and Andrew Nevins

Rules, Constraints, and Phonological Phenomena

This page intentionally left blank

Rules, Constraints, and Phonological Phenomena

Edited by

BERT VAUX AND ANDREW NEVINS

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published by Oxford University Press 2008

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd., King's Lynn, Norfolk

ISBN 978–0–19–922651–1 (Hbk.)

1 3 5 7 9 10 8 6 4 2

Contents

Notes on the Contributors

Bert Vaux is a University Lecturer in Phonology and Morphology at the University of Cambridge and a Fellow in Linguistics at King's College, Cambridge. His research focuses on the structure and origins of the phonological component of the grammar, especially in the realms of psychophonology, historical linguistics, and sociolinguistics. He is editor with William Idsardi of the book series *Oxford Surveys in Generative Phonology*.

Andrew Nevins is an Assistant Professor of Linguistics at Harvard University, where he conducts typological and experimental research on vowel and consonant harmony, reduplication, and lexical representations.

Ellen Broselow is a Professor of Linguistics at Stonybrook University. She is currently an Editor of *Natural Language and Linguistic Theory*. Her research focuses on phonological theory, second language acquisition as it relates to phonological constraints, and the match between phonological representation and phonetic structure. Much of her work has been on the phonology and morphology of Semitic languages, particularly Arabic dialects, though she has worked on Bantu, Indonesian, and Native American languages as well.

Paul Kiparsky is a Professor of Linguistics at Stanford University. His research focuses on Lexical Phonology and Morphology, Stratal Optimality Theory, metrics, and diachronic linguistics.

John Frampton is an Associate Professor of Mathematics at Northeastern University. His linguistics research focuses on the derivation of linguistic structures and he has worked on extensively on reduplication, morphology, and syntax.

David Odden is Professor of Linguistics at Ohio State University. His primary area of research is descriptive linguistics and phonological theory, particularly concentrating on African languages as well as North Saami.

William J. Idsardi is an Associate Professor in the Department of Linguistics and a member of the Cognitive Neuroscience of Language Lab at the University of Maryland. His research is focused on Phonology and its connections to other components of grammar and mind (phonetics, morphology, syntax, poetry, acquisition, neurolinguistics and computational models) and he has published papers on speech perception, metrical structures, opacity and rule systems, phonological features, reduplication and serial order, statistics in language and computational complexity.

Eric Raimy is an Assistant Professor in the Department of English at the University of Wisconsin, Madison and a member of the Wisconsin Englishes Project. His research is focused on the role precedence plays in phonology, language acquisition, and computational linguistics. He has published papers on language acquisition, language games, and reduplication.

Charles Reiss is an Associate Professor of Linguistics at Concordia University in Montreal. He is interested in theoretical linguistics, cognitive science, and historical linguistics.

Abbreviations

LPM-OT	Lexical Phonology & Morphology OT
LP	Lexical Phonology
OT	Optimality Theory
OT-CC	OT with candidate chains
OCP	Obligatory Contour Principle
C	consonant
V	vowel
ROTB	Richness of the Base
CM	Comparative Markedness
TCOT	OT with Targeted Constraints
OP	Optimal Paradigms
SPE	The Sound Pattern of English (Chomsky and Halle 1968)
BR	base-reduplicant
IO	input-output
OO	output-output
RED	reduplicant
CT	Correspondence Theory
GP	generative phonology
RBP	rule-based phonology
IPA	International Phonetic Association
DY	Duke of York
UG	Universal Grammar
PLD	primary linguistic data
H	high tone
L	low tone
TBU	tone-bearing unit

Symbols

☞	Points to winning candidate in an OT tableau.
❁	Marks a "flower" or "sympathetic" candidate, to which the winning candidate output must be faithful. Also employed (cf. Odden figure (11)) to mark a Sympathetic constraint, whose function is to identify faithfulness constraints violated by a given output candidate *vis-à-vis* the flower candidate.
❋	Marks a "selector" constraint used to identify a flower candidate.
💣	Flags a candidate that wins in the tableau but is not the actual output in reality.
☜	Indicates a candidate that wins in reality, but not in the tableau.
$\vee$	Conjoins two constraints (cf. Odden figure (41)).
✿	Marks an incorrectly selected sympathetic candidate.
☆	Used to mark a constraint crucial for the argument in Odden's (87).
$\aleph_0$	Countably infinite cardinality.

This page intentionally left blank

1

Introduction: The Division of Labor between Rules, Representations, and Constraints in Phonological Theory[1]

ANDREW NEVINS AND BERT VAUX

1.1 Introduction

Theoretical phonologists in the rule-based tradition (RBP)[2] represented by Chomsky and Halle (1968) and the Optimality tradition represented by Prince and Smolensky (1993/2004) have developed highly articulated models of the phonological component of the grammar that focus on three objects of inquiry: (i) phonological representations, (ii) processes that affect these representations, and (iii) constraints that delimit the scope of these representations and processes. Though in both frameworks certain phenomena have emerged as central, such as opacity, markedness and (under)specification, and morphology-phonology interactions (e.g. cyclicity, paradigmatic effects, exceptionality, and over-/underapplication), the ways in which these phenomena are understood in each framework can be quite different and often incompatible. It is thus important when constructing a phonological analysis to consider the relevant phenomena from both rule-based and constraint-based perspectives, as each has its own insights to offer: rule-based models, for example, enable us to make sense of opacity and cyclic effects in ways not possible with monostratal OT (leading in some cases to the development of

[1] Thanks to Eric Baković, Ricardo Bermúdez-Otero, Andrea Calabrese, Bill Idsardi, Greg Iverson, and Charles Reiss for comments on earlier drafts of this chapter.

[2] We follow McCarthy (2006) in referring to the Chomsky and Halle (1968) tradition as RBP, though in principle this conceals a number of additional distinctions, such as the fact that (i) RBP can employ constraints and/or be monostratal, and (ii) OT can be serial and/or derivational. As the fact that RBP employs rules and OT does not is an essential difference on which everyone can agree, we adopt the label in this form here.

stratal architectures of constraint evaluation; cf. Chapter 6 by Kiparsky) while constraint-based models have changed our thinking about apparent ordering paradoxes and the teleology of phonological phenomena (look-ahead effects, overapplication, and defect-driven phonological operations).

Because both rule-based and constraint-based frameworks can enrich our understanding of phonological phenomena and of the overarching structure of the phonological component, we have set out in the present volume to investigate the three above objects of inquiry using contributions by seasoned scholars versed in both frameworks, who are able to conduct informed comparison using the conceptual and analytical tools provided by both models.

1.2 Rules and temporal sequencing

Starting with the second object of inquiry above, we ask: what is the role of rules in the phonological component? This question encapsulates two central issues in current cognitive scientific research: the role of temporal sequencing in processing and storage (serial vs. parallel derivation), and the form in which humans extract and encode generalizations about their perceptual worlds (rules vs. activation values/constraint rankings). Directly relevant to these two issues is Gallistel's (2006) observation that theories of cognition are generally of two sorts:

(1) Theories of cognition

i. "Learning that" theories, in which hypotheses about the world are encoded in the form of generalizations or rules.

ii. "Learning to" theories, in which no hypothesis formation occurs, but rather incoming stimuli alter settings of universal elements (e.g. neurons or constraints).

Gallistel likens type (i) theories to Turing machines (i.e. generative procedures), and type (ii) theories to recurrent switching networks. Translated to a linguistic context, rule-based models are of type (i) and constraint-based models are of type (ii), since their central mechanism involves automatic rearrangement of a predetermined set of elements (in this case, constraints) in response to sets of incoming stimuli.

Given this general classification of models of cognition, the question arises in the case of the human phonological component whether these two types of models have empirical differences: can we identify any predictive differences between the two, or are they extensionally equivalent? While Gallistel demonstrates that theories of type (ii), unlike theories of type (i), are unable to generate cognitive processes such as dead reckoning and temporal learning, do they differ in the linguistic sphere? Nevins (2007) observes that researchers across various disciplines from propositional logic to theoretical linguistics believe that rules and constraints are always logically intertranslatable. Mohanan (2000: 145ff.) for instance asserts that a rule p → q is equivalent to a constraint ¬(p & ¬q), and McCarthy (1998) states that an ordered rule pair of the sort in (2) is equivalent to a partially ranked constraint set (3).

(2) A focus counterfeeding rule pair and its equivalent constraint set

Underlying Representation		/ABC#/
Rule 1	D → E /A___	—
Rule 2	B → D /___C	ADC#
Surface Representation		[ADC]

(3) Equivalent rankings
Faith(B ↛ E) >> {*AD >> Faith(D ↛ E)}, {*BC >> Faith(B ↛ D)}

Many researchers might take the putative equivalence of rule/constraint sets of the sort in (2)–(3) to support a model which uses only constraints: since constraints are needed independently to capture conspiracies, they reason, why add rules into the mix when their functions can be captured equally well by constraints?

In response, Calabrese (2005) argues that even if rule and constraint systems are intertranslatable, this does not mean that a given phonological model must have only rules or only constraints; logical equivalence does not imply equivalent explanatory adequacy for a given phonological phenomenon. According to Calabrese, an ideal phonological theory should contain both constraints and rules, each serving a distinct function: constraints are instructions to *avoid* a given configuration; rules are instructions to *create* a given configuration.

Employing both rules and constraints allows for a principled distinction between what is a "conspiracy", e.g. the ban on three consecutive consonants (*CCC; cf. Kisseberth 1970), and what is not, e.g. the ban on open syllables ending with [a] (*a]$_{Syll}$; McCarthy 2000). Constraints represent configurations that are systematically avoided within and across languages by means of various processes; *CCC for instance could be avoided by epenthesis, deletion, or fusion, and represents a configuration that in language after language is eschewed however possible. Similarly, processes like hiatus resolution can be modeled as *VV constraints, which trigger a variety of repairs (Casali 1997).

By contrast, McCarthy's *a]$_{Syll}$ is designed to impel open-syllable [a]-raising in Bedouin Arabic. However, [a] in an open syllable is by no means a cross-linguistically avoided configuration, nor is it associated with a variety of repairs. Bedouin [a]-raising is therefore better modeled as an instruction to create a new configuration from an old one.

While many theorists would agree that the division of labor between rules and constraints parceled out in terms of avoiding vs. creating configurations is reasonable, important questions remain for the learner faced with analyzing a given phenomenon as one or the other. This aspect of a model with both constraints and rules deserves further research. In addition the rules/constraints distinction may be compared on a number of other grounds, many of which suggest that the two are not in fact extensionally equivalent. As is discussed in greater detail in Chapter 2 of this volume and in Vaux (2007), existing incarnations of the two classes of theories have a number of predictive differences. We turn to a few of these.

Rule-based phonology (RBP), when coupled with a learning theory that employs information-theoretic learning (Gallistel 2003; Tenenbaum and Griffiths 2001, as proposed in Vaux 2007), or with a parameter-setting approach to phonological rules (e.g. Dresher and Kaye 1990; Dresher 1999), allows for a quick formation of rules upon exposure to limited data, consistent with LouAnn Gerken's generalization (pers. comm.) that infants often form linguistic hypotheses following exposure to as few as three data points and with Tenenbaum's work on human inductive leaps from as little as one data point (Tenenbaum 1999; Tenenbaum and Griffiths 2001). In contrast, existing OT learning theories (Recursive Constraint Demotion (Tesar and Smolensky 2000) and the Gradual Learning Algorithm (Boersma and Hayes 2001) require large numbers of {data exposure + learning} iterations to acquire the equivalent of an average phonological rule. The jury is not yet entirely in on which of these predictions is to be preferred, as the state of our understanding of the acquisition of phonological processes is still fairly limited; the key for our present purposes is that the two theories differ with regard to the predictions they make.

Nevins and Endress (2007) conducted an experiment in which participants were presented with an ambiguous rule involving trisyllabic sequences of nonce syllables: 123 → 321 (e.g. *ka.lei.bo* → *bo.lei.ka*). This transformation is compatible with at least four hypotheses:

(4) a. Invert the order of syllables
 b. Exchange the first and last syllable
 c. Exchange the final and antepenultimate syllable
 d. Exchange every other syllable (i.e. σ_j with σ_{j+2})

These hypotheses differ in the instances or kinds of positions they explicitly name, e.g. first, last, antepenult. In principle, upon hearing 123 → 321, participants might have chosen any of the hypotheses in (4), all of which account for the data. Nevins and Endress asked which of these hypotheses (or any of them) participants would select after exposure to just twenty-five trials. These precedence-modifying transformations are well-modeled in terms of a transposition rule but would be difficult to express concisely in terms of violations of surface constraints. Importantly, these four hypotheses all diverge on their predictions for an input string in which there are tetrasyllabic inputs, as shown for the hypotheses in (4) in their respective order:

(5) a. Invert the order of syllables: 1234 → 4321
b. Exchange the first and last syllable 1234 → 4231
c. Exchange the final and antepenultimate syllable 1234 → 1432
d. Exchange every other syllable (i.e. σ_j with σ_{j+2}) 1234 → 3412

The hypotheses in (4c) and (4d) are unexpected based on the existing typology of language games: there are no extant precedence-modifying language games that refer to "penultimate" or "every other" syllable. Nevins and Endress found that (5a) and (5b) were massively preferred by the participants, suggesting both that the participants were able to formulate a generalized version of this transformation with only a short number of trials and moreover that this hypothesis formation was constrained by representational primitives derived from the research into what is a possible rule.

Another point on which classic RBP and classic OT differ is the well-worn phenomenon of opacity (cf. Baković 2007 for a typology of opacity effects). RBP predicts the existence of opaque rule orderings (assuming that opaque generalizations are inferrable by language learners from primary linguistic data); classic OT predicts the absence of a subset of these, namely counterbleeding and environment counterfeeding systems (McCarthy 2000; see Collischonn 2007 for discussion of such a counterfeeding system in European Portuguese, and Wilson 2006 for a different counterfeeding effect problematically predicted to exist by OT-CC, "counterfeeding from the past"). And when a data set is compatible with both an opaque and a transparent interpretation, many forms of OT predict that the learner will opt for the transparent interpretation, whereas RBP allows for learners to choose the opaque option (Vaux 2006). In addition, RBP predicts the existence of iterative optional rules such as French schwa deletion (Dell 1980); classic OT predicts the absence of such processes in human languages (see Chapter 2 of this volume for further discussion).

RBP and OT differ computationally as well. Whereas RBP is computationally tractable and easily rendered in terms of finite state automata (Howard 1972; Johnson 1972; Idsardi 2007), Karttunen 1998 shows that OT is not computationally tractable unless one introduces the additional restriction that no constraints are evaluated gradiently.[3] Idsardi (2006) shows moreover that classic OT (unlike RBP) is NP-complete, thus requiring an extremely demanding class of computational complexity (though see Kornai 2006 for critique of this generalization).

In sum, there are reasons to believe that one cannot simply adopt an exclusively rule-based or constraint-based model without formal and empirical consequences. These consequences are particularly visible in the context of our discussion of rules: OT by definition does not allow for the encapsulation of linguistic hypotheses in the form of rules, whereas RBP does; RBP explicitly sequences the application of rules in time and allows for each intermediate rule and representation to be targeted by external processes (language games, rhyme schemes, speech errors, etc.), whereas classic OT states that intermediate stages cannot be accessed.

Interestingly, McCarthy (2006) and Pater (2007*b*) have recently conceded that potentially unbounded serial computation of the sort employed in RBP needs to countenanced within an OT framework as well. If we allow for staged processing of this sort (in either RBP or OT), several important questions arise, each of which is addressed in multiple chapters of the present volume:

(6) Central questions of ordering in phonology

a. Is rule ordering a necessary part of the phonological computation?

b. Are distinct levels or strata of phonological computation needed?

c. Are rules iterative, and if so what are the consequences?

Let us consider each of these in turn.

1.2.1 *Rule ordering*

The first question in (4) is raised for example by the opaque interaction between lax-vowel harmony and pre-fricative tensing in Canadian French, as analyzed by Poliquin (2006).

[3] McCarthy (2003*b*), for independent reasons, argues that gradient constraints can and should be replaced by categorically evaluated versions.

(7) Canadian French vowel harmony

a.	midi, *mɪdi, *mɪdɪ 'noon'	(vowels are underlyingly tense)
b.	*fi.lip, fi.lɪp 'Phillipe'	(Laxing: vowels in final closed syllables are laxed)
c.	fi.lɪp ∼ fɪ.lɪp 'Phillipe'	(Harmony: preceding high vowels optionally harmonize)
d.	mi.siv ∼ mɪ.siv 'letter'	(Tensing: pre-fricative tensing of vowels in final syllables)
e.	rule ordering: Laxing ⇒ Harmony ⇒ Tensing	

The second form in (7d) shows apparent overapplication of lax-vowel harmony: the pre-final syllable is open and therefore should not have a lax vowel. If Tensing were to precede Harmony, the former would bleed the latter, and attested forms like [mɪ.siv] would not occur. This is thus an instance of counterbleeding opacity, which poses a larger cognitive and functional problem: why should a final vowel be laxed only to be unlaxed at a later stage in the derivation? Odden's discussion of Duke of York (DY) derivations in Chapter 3 merits special comment here, in light of the general goals of the present volume. In DY derivations, a structure may be changed by one rule and subsequently restored by a later rule, with no "look-ahead" or "look-behind" to mediate in this process (cf. Pullum 1976; McCarthy 2003*a*; and Chapter 2 of this volume). DY derivations pose a problem for functionalist and monostratal constraint-based models, as it is unclear in these why a process (if triggered by a functional constraint) would apply if it is only going to be subsequently undone (indicating that the original functional constraint is in fact trumped by another constraint).

Pullum (1976) suggests that natural languages do not employ DY derivations, and McCarthy (2003) argues at great length that feeding (or, more precisely, non-vacuous non-paradigmatic) DY derivations do not exist, and that Sympathy Theory excludes just such derivations.

Odden shows that DY derivations actually *can* be generated by an OT grammar, and *must* be generated by any grammar to account for phenomena in the Bantu languages of Tanzania,[4] so that there is little reason to prefer constraints over rules based on considerations of expressive power. He demonstrates in addition that Sympathy Theory can generate transitivity violations (where process A precedes B, and B precedes C, but C precedes A), while rule ordering cannot. He shows moreover that two-level constraints (a possible alternative to Sympathy) are not rich enough to express transfer of tone from a moraic

[4] Bermúdez-Otero (2006*a*) demonstrates that another such gambit occurs in Catalan.

nasal in the process of tone deletion known as Meussen's Rule. From these three sets of arguments, Odden concludes that rule ordering is at least as good a model as any of its competitors.

As phonology increases its integration with the cognitive sciences, it is important to augment internal evidence of the sort adduced by Odden (following in the footsteps of early work by Voegelin and Swadesh 1935; Bloomfield 1939; Wells 1949; Chomsky 1951, etc.) with relevant evidence from other domains. As den Ouden (2001: 56) points out, the debate between OT and DP should take into account, for example, the results of psycholinguistic studies on temporal processing of language. Such studies have in fact found evidence for intermediate stages of representation with regard to English d → z vs. z → s as in *decide:decisive* vs. *permit:permissive* (Chomsky and Halle 1968: 229; Anisfeld 1969), spoken word recognition (Kolinsky 1994), and Cuna syllable reversal vs. degemination and devoicing (Sherzer 1970; for further discussion of language games accessing different derivational stages, see Churma 1985: 89–90).

The principles of staged phonological computation are re-examined in light of new phenomena and new predictions in the chapters by Paul Kiparsky, David Odden, and Bert Vaux. All three authors raise important arguments for sequenced computation, Odden and Vaux in terms of rules and Kiparsky in terms of ordered levels of constraint evaluation. These three authors embrace extrinsic ordering, with Odden and Vaux hypothesizing an extrinsic ordering for rules, and Kiparsky hypothesizing an extrinsic ordering for constraint evaluation within and across levels of representation. (All forms of OT use extrinsic ranking of constraints.)

The chapters by Frampton and Reiss also examine rules, but from a different angle. They formulate particular hypotheses about the nature of structural descriptions, raising questions about the degree to which phonology is dependent on formal logic, and the degree to which it is defect-driven. Frampton envisions an architecture that avoids enriched structural description by placing limitations on what the output of a structural change may yield, whereas Reiss outlines an architecture in which quantification plays an important role in what makes a possible rule.

1.2.2 *Levels and cycles*

Returning to our organizing questions concerning rules and their application in (6), question (b) asks whether distinct levels of phonological computation are needed. This touches not only on the classic level issue of Lexical Phonology (Kiparsky 1982) and Lexical Phonology and Morphology in OT

(LPM-OT, Kiparsky 2000), but also on the issue of cyclicity. Monostratal optimality-theoretic models typically attempt to deal with both level and cycle effects via Output-Output constraints (cf. e.g. McCarthy 2003*a*), whereas Stratal OT models treat at least a subset of these using a number of levels (typically Stem, Word, and Phrase) characterized by different rankings of the same set of constraints. Kiparsky 2000 argues persuasively that a stratal version of OT is superior to a monostratal version in dealing with opacity effects and in correlating conditions of process application with morphological structure.

In his contribution to the present volume, Kiparsky builds on this stratal model by analyzing three phonological phenomena that have played central roles in recent phonological theory: foot-structure well-formedness, level ordering, and the division of labor between non-phonemic enhancement and phonological contrast. He focuses on a difference between West Swedish (spoken in Sweden), where words containing only light syllables are not allowed, and Fenno-Swedish, spoken in Finland, where words such as *daga* are allowed. In the most basic pattern, open syllables are lengthened under stress. This reflects a tendency for feet to be binary, and also the effect of a stress-to-weight principle. Kiparsky's idea is that dialect variation is the result of promotion of a markedness constraint in the postlexical stratum of phonological computation. In dialects in which lengthening occurs for CVC words as well, this is due to promotion of the markedness constraint requiring consonant extrametricality.

In addition, Kiparsky notes that in West Swedish CVVC syllables behave specially in inducing vowel shortening, only when the final consonant is the first half of a geminate. In this case, the consonant must be analyzed as weight-bearing, and count towards a superheavy syllable. Notably, all dialects of Swedish lengthen a coda consonant when the stressed vowel is short.

Finally, Kiparsky analyzes the process of fortition, distinct from gemination, which can apply either postconsonantally (e.g. *ventta*) or postvocalically (e.g. *ruuppa*). The traditional substratum-based view is that Fenno-Swedish dialect variation (e.g. rural *riita* vs. urban *riitta*) is due to perception of the phonetically lengthened Swedish stop as either phonologically short or long. According to Kiparsky, this choice correlates with whether superheavy syllables were already contrastive in the dialect, and constitutes an anti-neutralization constraint. (Perhaps another way of viewing it is that postvocalic fortition is an enhancement effect between light and heavy syllables, in the sense of Stevens, Keyser, and Kawasaki (1986) and Keyser and Stevens (2006), who note that enhancement never recruits contrastive features.) Finally, Kiparsky uses stratal OT to derive the distinction between lexical and function words in Helsinki and Turku Swedish: function words enter the derivation as words, skipping the initial stem level where lengthening applies.

An interesting challenge for LPM-OT and other models that have limited numbers of stages of computation independent of morpheme-count is posed by phenomena that appear to require a discrete level for each morpheme in a word. A typical case occurs in Qashgar Uyghur (Orgun 1996*a*), where a non-word-final low vowel becomes high if it is the last vowel in a morpheme and in an open syllable (Raising, (6a)) and high vowels delete when flanked by identical consonants and between two open syllables (Elision, (6b)).[5]

(8) Qashgar Uyghur (Orgun 1996*a*)

a. Raising

kala 'cow'	kaliʁa	'cow-dative'
qazan 'pot'	qazɨni	'pot-possessive'
bala 'child'	balɨsi	'child-possessive'
ana 'mother'	anɨlar	'mother-plural'

b. Elision

/qazan-i-ni/	→	[qazɨnni]	'pot-possessive-accusative'
/bala-lar-i/	→	[ballɨri]	'child-plural-possessive'

In a number of possible multistratal models (though not LPM-OT in its current form), the number of phonological cycles may be equal to the number of affixation operations generating the morphological structure of the form: /bala-lar/ 'child-plural' for example undergoes one cycle, /bala-lar-i/ 'child-plural-possessive' undergoes two cycles, and /bala-lar-i-ni/ 'child-plural-possessive-accusative' undergoes three cycles. The fact that each affix triggers a new sequence of cyclic rule application can be inferred from the multiplicity of interactions between Raising and Elision in (9).

(9) Interactions between Raising and Elision

a.	/qazan-i-ni/	→	qazɨnni (*qazanni)	*elision counterbleeds raising; raising unexpected in surface closed syllable*
b.	/bala-lar/	→	balɨlar (*ballar)	*raising counterfeeds elision*
c.	/bala-lar-i/	→	ballɨri (*balɨlɨri)	*raising feeds elision*

The facts in (9) can be straightforwardly accounted for in RBP by assuming that both Raising and Elision are cyclic, with the former ordered before the latter, yielding derivations of the sort in (10).

[5] The accusative suffix has two allomorphs, /-i/ after consonant-final stems and /-ni/ after vowel-final stems. The same relationship holds for possessive /-i/ ∼ /-si/.

(10) Sample derivations

		a. 'pot-poss-acc'	**b. 'child-pl-acc'**	**c. 'child-pl'**
UR		/qazan-i-ni/	/bala-lar-i/	/bala-lar/
Cycle 1	input	qazan-i	bala-lar	bala-lar
	Raising	qazɨni	balɨlar	balɨlar
	Elision	—	—	—
Cycle 2	input	qazɨni-ni	balɨlar-i	
	Raising	—	balɨlɨri	
	Elision	qazɨnni	ballɨri	
SR		[qazɨnni]	[ballɨri]	[balɨlar]

Orgun observes that in order to account for the difference between [ballɨri] and [balɨlar], the outputs of the two rightmost derivations in (10), the number of phonological cycles must crucially depend on the number of suffixes, precisely what is not allowed in three-level models such as LPM-OT (cf. also Harmonic Phonology (Goldsmith 1993), another three-level model). An open possibility is that LPM-OT, and in fact, level-ordered models more generally, could allow multiple cycles *within a level*, where each level is distinct in terms of its rules or constraints and their ordering.[6]

To illustrate the problems the Uyghur data cause for LPM-OT, let us first postulate the informal constraints in (11) and rankings in (12) to mimic the effects of Raising and Elision.

(11) Informal constraints for Uyghur raising and elision

a.	*a	no non-final [a] in morpheme-final open syllables
b.	$*C_1V_{hi}C_1$	no high vowels between identical consonants in open syllables
c.	MaxHi	don't delete [high] specifications
d.	MaxLo	don't delete [low] specifications
e.	MaxV	don't delete vowels

(12) Rankings for Uyghur raising and elision

a.	*a >> MaxV >> MaxLo	yields vowel raising rather than deletion as resolution of (11a)
b.	$*C_1V_{hi}C_1$ >> Linearity, MaxHi >> MaxV	yields vowel deletion rather than metathesis or lowering as resolution of (11b)

6 Ricardo Bermúdez-Otero (pers. comm.) maintains that both Kiparsky's and his versions of multistratal OT have internal cyclicity at the stem level (whether directly or epiphenomenally), so that there can be more than one stem-level cycle (cf. Bermúdez-Otero and McMahon 2006: §3.4 and Bermúdez-Otero 2007 for further discussion).

At what level(s) do the rankings in (12) hold? The Stem level is not relevant here, as all of the attested effects of raising and elision appear at the word level. By the same token, this ranking cannot hold for the Phrase level, as raising and elision do not appear in phrasal contexts. Thus, LPM-OT must have the rankings in (10) only in its Word-level ranking.

This ranking will treat the URs in (10) in the manner outlined in (13).

(13) Uyghur outcomes using ranking (12)

/qazan-i-ni/	*$C_1V_{hi}C_1$	*a	MaxHi	MaxV	MaxLo
qazanini	*!	*			
qazɨnini	*!				*
qazɨnni			*!	*	*
qazanni			*!	*	
💣 qaznini				*	*
/bala-lar-i/					
balalari		*!*			
balɨlari	*!	*			*
ballari		*!		*	*
ballɨri				*	**!
balalɨri		*!			*
balɨlɨri	*!				**
💣 balalri				*	*
/bala-lar/					
balalar		*!			
☞ balɨlar					*
ballar				*!	*

It should be clear from (13) that no ranking of the relevant constraints in (11) can generate the desired range of results. One reason for this is the counterbleeding interaction between raising and elision in *qazɨnni* (cf. 10a); as McCarthy (2003*a*) has demonstrated, counterbleeding interactions cannot be modeled in a conventional monostratal evaluation. Put briefly in slightly more formal terms, *qazɨnni* can never emerge as the winner because it is harmonically bounded by **qaznini*. By the same reasoning, the desired form *ballɨri* can never beat **balalri*, because the former is harmonically bounded by the latter.

One might consider invoking additional syllable contact constraints to target the -*zn*- in **qaznini* and the -*lr*- in **balalri*, but (*a*) there is no evidence for Uyghur making use of such constraints, and (*b*) positing and highly ranking such constraints would incorrectly generate **qazanni* and fail to deal with the fact that **qazanni* (and not just *qaznini*) harmonically bounds the desired form *qazɨnni*.

The derivations in (10) could be modeled by having a separate constraint ranking and evaluation process each time one adds on an affix (as Orgun 1996*a* does), but it should be clear that the correct derivations cannot be obtained if one is limited to Stem, Word, and Phrase levels of evaluation. To paraphrase Orgun (1996*a*), the crucial limitation of the Harmonic Phonology and LPM-OT approaches is that every form undergoes the same number of applications of phonology, regardless of the morphological structure. Output-Output constraints are also unable to produce the Uyghur system, as the attested effects correlate only with affix count, not with properties of base forms. Considering /qazan-i-ni/, for example, there is no obvious way to get an OO constraint to favor the desired *qazɨnni* over its competitor **qazanni*.[7]

The Uyghur data thus provide strong evidence for both cyclic application and ordering of phonological processes. Bermúdez-Otero (2007) provides further evidence of the need for traditional cyclicity, even within an OT framework, based on the existence of masked bases and absent bases in non-canonical paradigms. Future research will yield light on whether such effects can perhaps be modeled with recurrent passes through a word-level stratum.

1.2.3 *Iterative application*

The final aspect of ordering that we consider in this volume is iteration (question (6c) above), a robust linguistic phenomenon first examined systematically in a phonological context by Howard (1972). Of particular interest for our purposes are processes that are both iterative and optional. One can conceive

[7] Related to this general limitation, Bobaljik (1998) observes that using OO constraints to account for cyclic effects makes incorrect predictions with regard to epenthesis in Itelmen, where epenthesis behaves cyclically in verbs (which have no bare base forms) and non-cyclically in nouns (which actually do have bare base forms).

of three ways in which iteration might surface in such cases, summarized in (14).

(14) Typology of interaction between iterativity and optionality

a. **all or nothing application**: if one target segment undergoes the process, then all of the target segments do; if one target segment doesn't undergo the process, then none of the target segments does.

b. **local application:** each target is evaluated independently of the others.

c. **locality-respecting application:** apply successively to each target, but if a given target does not undergo the rule, then cease scanning through the remainder of the relevant phonological domain for further targets.

One might say that processes of type (a) are evaluated globally (at the word level), and those of types (b) and (c) are evaluated locally (hence the name "local optionality" coined for type (b) by Riggle and Wilson (2006) for what Vaux (2003) originally labelled "sequential iterative optionality").

An example of type (14a) identified by Vaux (2003) is Warao labial voicing (Howard 1972: 87). Rules of type (14b) include English flapping, Dominican Spanish s-insertion, and French schwa deletion. Type (14c) is attested in Brazilian Portuguese (José Olímpio Magalhães, pers. comm.), where the noun *mexeríca* 'tangerine' for instance has the possible outputs in (15):

(15) Outputs of *mexeríca*

a. mexeríca

b. mexiríca

c. mixiríca

d. *mixeríca

We can straightforwardly generate the forms in (15) by postulating an optional rule of pre-tonic vowel reduction that moves leftward beginning with the stressed syllable. It may optionally apply successively to each potential target, but if it fails to apply once, the rule continues no further.

Yet another interesting case of "no skipping" is found in Shwayder's (2007) experimental study of Icelandic umlaut and reduction. Umlaut converts *a* → *ö* before a suffix containing *u*. Vowel reduction can in turn convert a non-initial umlauted syllable into *u*. In an experimental nonce-word study, Shwayder found that speakers would convert trisyllabic inputs like *ramanað+um* into *römönöðum*, *römönuðum*, or *römunuðum*, but not *römunöðum*. In other words, it is optionally possible to reduce the final unstressed umlaut syllable or all unstressed umlauted syllables, but it is not possible to skip the final syllable and reduce only the medial syllable. This too can be modeled as a right-to-left optional application of reduction, where if it fails to apply once, the rule continues no further.

Finally, a similar phenomenon occurs (but applying from left to right) in the penult-stress language Shimakonde (Liphola 2001:170 and Odden's chapter in this volume), in which unstressed mid vowels may be optionally reduced to [a], e.g. *li-kolomoódi* "cough" ∼ *li-kalomoódi*, *li-kalamoódi*, **li-kolamoódi*; thus, either both pre-tonic or the initial vowel may reduce, but the initial vowel may not be skipped in reduction with reduction subsequently applying to the second syllable.

In Chapter 2, Vaux identifies optionality of types (14b) and (14c) as core challenges for classic OT. In Howard's (1972) version of RBP, in which rule foci are sequentially evaluated with a pointer, it is possible to capture rules of type (14b), all of which can optionally apply at each locus where the structural description is met. In classic OT, where constraints are wholesale evaluations of a word's violations of them, there is no possibility, even with the devices that allow for constraint ties, for a single constraint to be optionally violated at some but not other locations within a word. In recent work, Riggle and Wilson (2006) and Pater (2007*b*) acknowledge this problem for classic OT and present divergent solutions.

John Frampton in Chapter 7 also examines iteration, but from a somewhat different perspective. He proposes extensions to the mechanism of iterative rule application, including avoidance constraints that limit the application of rules in certain instances. His study furthermore contains the idea that rule application is defect-driven in the sense that a rule applies in order to remove illicit or illegible elements from representations. Importantly, these well-formedness conditions need not be exclusively output-based in Frampton's model. He provides examples of footing and syllabification in a variety of languages, showing the interaction of repair rules and constraints for iterative sequences.

1.3 The Representational Vocabulary for Segmental and Suprasegmental Phenomena

Returning to the central objects of phonological inquiry that we identified at the beginning of this chapter, we would now like to move on to the first object we identified, phonological representations. In this realm we focus on the questions of the non-featural content of phonological representations at the segmental and suprasegmental levels. How is linear order encoded? How are variations in phonological weight within and across grammars best modeled?

William Idsardi and Eric Raimy (Chapter 5) address the level of segmental representation, where reduplication, metathesis, infixation, and other operations that alter precedence relations take place. Their work represents an effort to explore the cognitive delimitation of sequences, raising the question of shared cognitive structures between phonology and the procedural nature of musical copying. Idsardi and Raimy compare a constrained model of base-reduplicant relations, the Multiprecedence and Linearization model of Raimy (2000*a*), with Correspondence Theory, concluding that the latter allows an excessive number of mappings between base and reduplicant. They suggest that reduplicative structures may be adopted by learners for reasons of economy of description, and point out suggestive experimental results from musical cognition. In addition, they point out that the model provides a natural metric for comparing economy of computation. Idsardi and Raimy suggest that the model may be extendable to providing a formal description of deletion and gemination, and show typologically odd and unnatural reduplication patterns that are predicted by competing models.

The levels of suprasegmental representation receive significant attention in this volume as well, with important representational distinctions being drawn between epenthetic and lexical syllables, for example, in Chapter 4 by Ellen Broselow. Paul Kiparsky in Chapter 6 addresses representational distinctions made in terms of syllabic quantity and the contrast afforded by these representations.

Broselow focuses on the fact that epenthetic vowels are often invisible for stress assignment. However, rather than analyzing this as the result of the relative order of stress assignment *vis-à-vis* epenthesis, Broselow presents a model in which two grammatical forces encoded as constraints militate against the stressing of epenthetic vowels: the role of metrical prominence in lexical retrieval, and the role of distinct stem shapes in inflectional distinctions.

The first, exemplified by Selayarese, Winnebago, and Northern Kyungsan Korean, is to avoid placing non-underlying material in prominent positions. In Selayarese, the loanword *balábasa* shows that the penultimate-stress pattern ignores the final epenthetic vowel, but in the loanword *solodére*, in which the second and last vowel are epenthetic, stress seems to take the final vowel into consideration in deriving penultimate stress. The generalization then seems to be that prominent feet should not contain epenthetic vowels where possible. The intuition is thus that prominent foot structure in loanwords—perhaps because of its importance for lexical access and recovery of the original loanword—should stick to material from the source language.

In the second set of phenomena, exemplified by Iraqi Arabic, first- and second-person verbal forms and third-person verbal forms show distinct stress patterns on their stem throughout the perfective conjugation. Traditionally, this is seen as a consequence of third-person forms being vowel-initial, while first- and second-person forms are consonant-initial. However, additional phenomena such as vowel shortening and glide retention seem to conspire in rendering stems with [+Participant] and [-Participant] person features distinct. The interesting conclusion here is thus that a variety of phonological phenomena may operate with the teleology of separating morphosyntactically distinct forms.

1.4 The Role of Constraints in Delimiting Grammars

Turning to the third and final object of inquiry we identified at the beginning of this chapter, we ask in this volume what is the role of (in)violable constraints in the phonological component. It is important to emphasize that investigation of the role of constraints in the phonology is in principle independent of the question of serial vs. parallel computation, and that it is logically coherent to ask questions about the groundedness of phonological representations and processes within a framework of non-violable constraints, as has indeed been done in the works of Kean (1975), Kawasaki (1982), Calabrese (1988, 1995), and Archangeli and Pulleyblank (1994), to mention only a handful of fairly recent representative research.

With this caveat about constraints as an independent architectural choice in mind, Kiparsky (Chapter 6) examines whether constraint permutations insightfully model cross-linguistic phonological variation, and Idsardi and Raimy consider in Chapter 5 whether Correspondence constraints optimally generate over- and underapplication effects. Vaux (Chapter 2) investigates whether a theory that uses only constraints is necessary and sufficient to account for the attested range of phonological phenomena, and Reiss (Chapter 8) considers whether the phonology should use constraints at all.

Reiss proposes that a more restrictive model of UG comes not from constraints on what can be a possible grammar, but from an exhaustive list of what are the possible primitives and operations on them. Thus, while generative phonology has often appealed to inviolable constraints (such as the

Obligatory Contour Principle, the Avoidance constraints of Idsardi (1992), the co-occurrence filters of Calabrese (1988), etc.), Reiss suggests that more explicit structural descriptions of structure-building rules can always yield a theory that lacks constraints of any kind. Reiss points out that while this may complicate the analysis of individual languages, it simplifies the theory of Universal Grammar. Reiss thus takes the strong position that constraints are unnecessary and undesirable as part of a model of phonological computation.

By contrast, Frampton (Chapter 7) argues that constraint-like "defects" in intermediate representations can provide triggering environments for rule application and that the sequence of rule application may be better understood in terms of (inviolable) constraints on various types of representations. Frampton's model is one in which constraints play an important role in "triggering" rule application, because certain structural descriptions are seen as defective and hence in need of elimination of repair.

1.5 Conclusions

In pursuing a complete understanding of the human phonological component, a number of sources of evidence constrain the space of models: learnability considerations, psycholinguistic and neurolinguistic evidence, and phonological phenomena in the languages of the world. The authors in this volume focus carefully on subsegmental, segmental, and suprasegmental phenomena in a diverse range of languages, with the aim of delimiting classes of compatible theories of phonological representations and computation.

Indeed, one common research strategy pursued by all of the contributors in this volume is to seek and analyze phonological phenomena with the specific goal of determining the consequences for models of cognition and language sound structure. In 1861, Charles Darwin remarked in a letter to a friend, "About thirty years ago there was much talk that geologists ought only to observe and not theorize; and I well remember some one saying that at this rate a man might as well go into a gravel-pit and count the pebbles and describe the colours. How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service!"

In this introduction, we have delimited the three fundamental objects of phonological inquiry addressed in this volume—rules, representations, and constraints—and we hope that, as the contributions in this volume are taken as a whole, certain unifying conclusions emerge concerning both the temporal sequencing of processes and the teleological nature of certain phonological phenomena. Some issues remain contentious in this volume, such as the roles played by cyclic application and the violable nature of constraints, but the contributions included here present us with strong arguments for each position that should enable us to emerge with a new consensus built upon broader and firmer ground than what was possible when approaching these problems from a single theoretical perspective.

2

Why the Phonological Component must be Serial and Rule-Based[1]

BERT VAUX

2.1 Introduction

This chapter provides general arguments for replacing Optimality Theory with a theory that employs ordered rules and derivations.

Between 1968 and 1993 the majority of phonologists worked within Rule-Based Phonology (RBP), whose central proposition is that the surface representation of a sequence of morphemes derives from their abstract underlying representations by the application of a series of ordered rules. The introduction of Optimality Theory (OT) in the early 1990s by McCarthy, Prince, and Smolensky has resulted in a drastic realignment of the field of phonology, in terms both of the questions that are being asked and of the ways in which these questions are being addressed. In canonical OT the underlying and surface representations are related by means of universal violable constraints, and the differences among languages are claimed to be due exclusively to differences in the rankings of these constraints.

The rapid acceptance of OT in North America, Europe, and East Asia could be argued to have been due in part to a dissatisfaction among phonologists with aspects of RBP such as its perceived lack of universality, the stipulative nature of its extrinsic rule orderings, the clumsiness of its inviolable constraints *vis-à-vis* the rankable and/or weightable constraints of OT (cf. Pater 2000; Chen-Main 2007), and its perceived failure to formalize satisfactorily the interaction between the rules and constraints it employs.[2] I have summarized

[1] A shorter version of this paper was read at the LSA Annual Meeting in Atlanta on January 3, 2003. Thanks to Morris Halle, Andrea Calabrese, Andrew Nevins, Justin Fitzpatrick, Laurie Karttunen, John Frampton, Bill Idsardi, Eric Raimy, Ellen Kaisse, Donca Steriade, Cheryl Zoll, Joe Pater, Fiorien Bonthuis, and the members of the MIT Phonology Circle for comments on earlier drafts.

[2] Thanks to Joe Pater (pers. comm.) for these last two points. He adds that "it's [an open question] whether OT satisfactorily formalized that interaction (i.e. by placing operations in a constraint-blind Gen)."

in (1) the claimed advantages of OT that I have been able to find in the literature.

(1) Arguments adduced in favor of OT over RBP

a. **New directions, new empirical results** (McCarthy and Prince 1993; McCarthy 2002*a*)

b. **Generality of scope** (The OT framework can be used for all components of the grammar, not just phonology and morphology; McCarthy 2002*a*.)

c. **Parsimony** (McCarthy 2002*a*: 243: "if a constraints-only theory is workable, then it is preferable [to a theory combining rules and constraints], all else being equal"; cf. Kager 1999: 187: OT is "conceptually superior" in that "we find that a rule-based analysis uses excessive machinery to achieve effects that an OT analysis attributes to a single interaction".)

d. **Direct incorporation of markedness** (Constraints actually produce cross-linguistic distributions and markedness rather than restating them; McCarthy and Prince 1993: 19; Eckman 2005.)

e. **Compatibility with connectionism** (Constraint systems of the OT type are attractive for implementation in terms of connectionist networks (Smolensky 1999; Dell et al. 1999; Seidenberg and MacDonald 1999). McCarthy (2002*a*) points out in his FAQ section that OT differs from connectionism in having strict domination, and Legendre *et al.* (2006) demonstrate that weighted constraints predict the existence of unattested systems that ranked constraints are unable to produce (cf. also Kiparsky 2005). It is interesting in this connection to note the increasing popularity of weighted constraints, which bring OT even closer to connectionism; cf. Hayes and MacEachern 1998; Mohanan 2000; Flemming 2001; Boersma and Hayes 2001 (though they maintain strict domination for parsing); Pater 2007*a*.)

f. **Factorial typology derives from free ranking** ("By assuming that *all* constraints have to be universal, OT severely restricts the degrees of freedom in model formulation in linguistics (one of the core problems of linguistic description)… OT furthermore offers a restrictive theory of linguistic variation: differences between languages can arise only a different rankings of universal principles in different languages" (Féry and Fanselow 2002). McCarthy 2002*a*: 113 claims moreover that OT provides a clearer picture than RBP does of typological "overkill" (also called the Too Many Solutions Problem) such as the absence of deletion as a repair for voiced coda obstruents: "OT, because of its inherently typological nature, calls attention to this problem [of overkill] and suggests where to look for a solution, based on harmonic bounding. In contrast, rule-based theories, at least in phonology, rarely address typological matters and offer no general solution to this problem.")

g. **Conspiracies** ("Compelling examples of homogeneity of target/heterogeneity of process tend to support constraint-based over rule-based theories" (McCarthy 1999*a*; cf. also McCarthy and Prince 1993: 4, Prince and Smolensky 1993: 1, etc.).)

h. **Morpheme Structure Constraints and the Duplication Problem** (Rules and phonotactics replicate each other; Kager 1999: 56 *inter alia*.)

i. **Problems with rules and levels** ("Besides the deus-ex-machina character of the level distinction itself, the additionally necessary manipulations indicate that this mode of phonological analysis [i.e. postulating word-internal levels without independent justification beyond the phenomenon under discussion] holds little promise" (Itō and Mester 2003*a*: *n*. 16 and associated text). "Compare the proliferation of strata in works like Halle and Mohanan (1985): four lexical strata, one of which includes a loop, plus the post-lexical stratum. This comes close to being a reductio ad absurdum of LP" (McCarthy 2004, handout on Stratal OT for Ling 730). Other problems cited with rules include their being unconstrained, arbitrary, language-specific, and requiring look-ahead and look-back power; rule systems are claimed to be unconstrained in their interactions and sometimes involve ordering paradoxes (cf. Anderson 1974) and pathologies (Prince and Smolensky 2002: 137).)

j. **Grammaticality judgements and gradient well-formedness** (Gradient well-formedness effects imply speaker knowledge of violated constraints; such effects are not modelable in RBP (Steriade 2000; cf. also Keller 1998; Hayes 2000; Coetzee 2004).)

k. **Back-copying/overapplication in reduplication** (e.g. oven → woven-way for a small percentage of Pig Latin speakers;[3] "Correspondence Theory is superior, empirically and conceptually, to serial derivational approaches. All serial theories are incapable of

[3] This is a subtype of the variety of Pig Latin that inserts -way after vowel-initial words (e.g. pig → ig-pay but ant → ant-way); see Vaux and Nevins 2003 for further details.

dealing with cases in which B copies (or, more neutrally, reflects) R" (McCarthy and Prince 1999: 290).)

l. "**Serial derivations are cognitively implausible**" (Orgun 1993; Sebregts 2001: 63; seriously undermined by OT-CC (McCarthy 2006, 2007).)

m. **Unification of description of individual languages with explanation of language typology** ("Joining of the individual and the universal, which OT accomplishes through ranking permutation, is probably the most important insight of the theory" (McCarthy 2002*a*: 1).)

n. **Learnability** ("If the constraint set is universal, this cuts down the [language learner's] analysis space considerably [as opposed to learning sets of ordered rules, especially extrinsic and opaque orderings]" (Zuraw 2004).)

o. **Separation of structural description and structural change** (Theories of structural descriptions and of structural changes are "loose and uninformative" and therefore "the locus of explanatory action is elsewhere" (Prince and Smolensky 2004: 4). Cf. "in a theory where phonological rules specify both context and change, as in SPE and much work following it, it is not possible to account for this asymmetry of [overkill] patterns except by stipulation" (Lombardi 2001: 13). Compare also Hayes (2004) on phonological acquisition: "within Optimality Theory, the learner must locate the Faithfulness constraint that must be ranked lower in order for underlying forms to be altered to fit the phonotactics. By way of contrast, earlier rule-based approaches require the learner to find both structural description and change for every alternation, with no help from phonotactic knowledge".)

Even in the earliest OT treatments these problems were mentioned only in passing;[4] I am not aware of any serious attempt by an OT supporter to explicitly examine or falsify an RBP analysis. This is not surprising, given that none of the points in (1) actually poses a legitimate problem for RBP.[5] McCarthy

[4] Cf. Prince and Smolensky's (2002: 22) critique of look-ahead power (cf. (1i)): "In I[mdlawn] T[achlhiyt] B[erber], however, as in many other languages, the availability of nuclei depends on the choice of onsets: an early step in the derivational constructive procedure, working on a low level in the structural hierarchy, depends on later steps that deal with the higher levels. Indeed, the higher level constraint is very much the more forceful. Technical solutions to this conundrum can be found in individual cases, Dell and Elmedlaoui's being a particularly clever one; but the theme will reappear persistently in every domain of prosody, defying a uniform treatment in constructionist terms."

[5] McCarthy's claim in (1f) that the OT treatment of overkill is superior to that of RBP, for instance, capriciously inverts the actual situation (no existing form of OT accounts successfully for overkill,

(1998: 4) states moreover in his discussion of serious problems posed for OT by opacity effects that "I will not attempt to respond to these critics here; the body of empirical and conceptual results directly attributable to OT makes a brief response both impossible and unnecessary." This variant of argument (1a) attempts to circumvent the scientific standards of accountability for falsification, which require that one formulate theories that are falsifiable and that one respond in good faith to subsequent falsifications, either by revising or abandoning one's theory, or by demonstrating that the attempted falsification was flawed. No amount of positive empirical or conceptual results is sufficient to override proof that a theory makes fundamentally incorrect predictions.

The parsimony argument in (1c) does not hold up, either. One can find just as many OT analyses that are forced by their framework to create byzantine appendages that are unnecessary in their RBP equivalent; a particularly clear example of this is Sympathy Theory as a response to the problem of opacity, as we will see later in this chapter.

Another of the advantages most often claimed for OT *vis-à-vis* RBP, factorial typology (1f), is nicely addressed by Kager 1999: 35: "the reranking approach would predict that any new grammar that arises from a reranking of any pair of constraints will precisely correlate with one of the world's languages. This prediction is based on the deeply naïve assumption that every possible ranking should be instantiated by some attested language. This is naïve, just as it is deeply naïve to expect that all logically possible permutations of genetic material in the human genome are actually attested in individual humans."

Space constraints prevent me from discussing the remaining points in (1) here; in what follows I focus on those that are mentioned most frequently in the OT literature. The issue in (1) that is most often cited involves conspiracies (1g); as McCarthy (1999*a*) puts it, "compelling examples of homogeneity of target/heterogeneity of process tend to support constraint-based over rule-based theories". Beyond that, work in OT to date, as exemplified by Prince and

whereas it is not a problem in RBP) and ignores the rich RBP tradition of Evolutionary Phonology (Ohala 1971, 1972, 1975, 1981, 2005; Ohala and Lorentz 1977; Chang, Plauché, and Ohala 2001; Hale and Reiss 2000; Vaux and Samuels 2004; Blevins 2004; Pycha *et al.* 2003, etc.) that provides an explicit account for overkill effects. McCarthy and Prince's claim concerning back-copying in reduplication (1k) is similarly false; Raimy's RBP model of reduplication (1999) can derive such effects, for example (see Raimy 2000*b* for pointed discussion). Baković 2007 asserts that cross-derivational feeding-on-environment in Lithuanian is a "teleological" problem for RBP, but acknowledges (p. 18) that Odden's rule-based analysis of the phenomenon (2005: 113–15) is "descriptively satisfactory". Since in my opinion there is no place for teleology in synchronic phonology, teleological objections of the sort raised by Baković (as opposed to substantive descriptive and predictive problems of the sort discussed in this chapter) are not a concern.

Smolensky's 1993 treatment of Berber syllabification, has consisted primarily of demonstrating that a constraint-based system can derive some of the same results as RBP.

Some have argued that the rise of OT was a classic paradigm shift in the Kuhnian sense. In fact, the shift from RBP to OT was quite different than the SPE revolution of the late 1960s: the latter generated a barrage of (ultimately productive) criticism, whereas the paradigm shift of 1993 was bloodless. The usual resistance and conservatism conveyed in top journals was circumvented by the development of the Rutgers Optimality Archive, which enabled younger phonologists to circulate developments of the new theory without being answerable to the objections of scholars more familiar with the body of facts that had led phonologists to espouse RBP's complex derivational machinery in the first place. To paraphrase Kiparsky 2000, once we look at entire phonological systems, not just toy examples of a few interacting constraints, we see that OT results in very serious loss of generalization. Chomsky (1967: 110) observes along similar lines that "to study the questions … in a serious way, one has to investigate a real language system with dozens (if not hundreds) of phonological rules, with complex ordering conditions among them determined on empirical grounds;… it is of no use to study a subsystem with three or four rules." This task has been carried out in hundreds of books and theses written in the RBP framework, but remains to be carried out for any language in an OT framework.[6]

Subsequent rediscovery of the facts that were already known in the RBP literature thanks to detailed investigations of this type has led in recent years to the reintroduction of core principles of RBP into OT, including levels (Kiparsky 2000; Rubach 2000), the cycle (Orgun 1993, 1996; Bermúdez-Otero 2007), constraints on underlying representations (Vaysman 2002; cf. Vaux 2005*a*), and most recently rule-like derivations (McCarthy 2006; Pater 2007*b*), but these modifications are not enough to save the theory, as I suggest in what follows.

Returning to the larger issue of the paradigm shift from RBP to OT, its end result has been that phonologists have moved to a different set of theory-internal issues without asking the bigger questions in (2) that should have been raised by a confrontation of the two perspectives:

[6] Michael Hammond's 1999 *The Phonology of English: A Prosodic Optimality-Theoretic Approach*, for example, covers only a tiny fragment of the phonological component of English. A search of amazon.com in May 2007 revealed only a single peer-reviewed, book-length OT treatment of the phonological component of a single language's grammar, Wheeler's 2005 treatment of Catalan. The closest I have been able to find on the Rutgers Optimality Archive is Picanco 2006.

(2) Central questions

i. What are the phenomena that a descriptively and explanatorily adequate theory of phonology must account for? (Compare the first sentence in Kager 1999: "the central goal of linguistic theory is to shed light on the core of grammatical principles that is common to all languages.")

ii. What phenomena do the two competing theories predict to be possible and impossible? (Compare McCarthy's dichotomy *"Can you do this one?" vs. √"What can/can't you do?" (1997: 12).) In other words, how exactly do the two theories differ?

I suggest that answering these questions leads to a specific theory of phonology that is serial and rule-based, along the classic lines set out in Kenstowicz (1994). In this chapter I assume a theory of this sort, building on the work of Halle and Vergnaud (1987); Halle and Marantz (1993); Halle (1995); Calabrese (1995, 2005); and Vaux (1998).[7] Space constraints prevent me from rehearsing the details of this theory here; I would like instead to focus on providing general arguments for abandoning RBP's primary competitor, Optimality Theory. The arguments fall into four basic categories, which I outline below. In doing so it is important to bear in mind that any reasonable and falsifiable theory will deal well with some phenomena and not so well with others. I therefore focus on *overarching* problems and *insurmountable* problems, rather than on small language-particular problems for which one theory happens to have a more efficient account than the other. I also pass over legitimate problems that have been identified in OT but happen not to be robustly instantiated cross-linguistically.[8]

2.1.1 *Central phenomena of human language*

The first major problem is that OT fails to account for several of the central phenomena of human language—i.e. those that occur in all or most known languages—which any adequate theory of phonology must be able to explain.

[7] Each of these works individually, as well all of them taken collectively, presents a highly detailed and coherent derivational, rule-based model of the phonological component. In light of this fact, it is unclear why OT supporters so frequently assert that proponents of RBP do not have an explicit theory of the synchronic and diachronic components of phonology. (Kirchner 2001: 428–9 for instance states that "proponents of the diachronic critique might meet this objection by presenting an explicit model of some aspect of the phonetics, or other domains giving rise to relevant functional principles, together with an explicit model of phonological acquisition and synchronic phonological grammar," wrongly implying that such explicit models do not exist.)

[8] A nice example is Wilson's (2003) demonstration that Classic OT allows for unattested non-local interactions of the sort "vowel epenthesis applies to a form with a final cluster except when there is a preceding [+nasal] feature anywhere in the word that is blocked from spreading to the right edge."

These phenomena include opacity, optionality, exceptionality, unnaturalness, and ineffability.

2.1.1.1 *Opacity* Opaque interactions between phonological processes occur in all known natural languages. This fact receives an elegant explanation in derivational models, wherein opacity is a straightforward product of process ordering. OT in turn is actually organized around a specific sort of opacity, namely constraints not being surface-true. As has been shown in great detail, though (cf. *inter alia* Idsardi 1997, 1998; Odden this volume; Kager 1999; Kiparsky 2000; McCarthy 2002), canonical OT encounters severe problems when dealing with the complex sorts of opacity that we actually find in natural languages, notably counterbleeding and environment counterfeeding (McCarthy 1997*b*, 2003*a*) and self-destructive feeding-on-environment (Baković 2007). Opacity created by iterative rules creates even more profound problems for OT, since proposed patches such as Sympathy, level ordering, and output-output constraints cannot be brought to bear (Wolfe 2000; Hyman and VanBik 2002). I discuss this problem in more detail in Section 2.3 below.

2.1.1.2 *Optionality* The second phenomenon to be accounted for is optionality. All languages contain numerous optional processes, a fact that is not predicted by the fundamental architecture of OT, as Kager 1999 and others have pointed out. OT mechanisms such as cophonologies and tied constraints fail (with the exception of Riggle and Wilson's (2006) local optionality scheme) to account for a variety of optionality effects such as sequential iterative optionality, as I detail in Section 2.5.

2.1.1.3 *Exceptionality and unnatural processes* Thirdly we must account for unnatural processes. A grammar arises from the confrontation of the human language acquisition device with the arbitrary linguistic data to which it is exposed. Since these data encode layers of historical change, the resulting phonological grammar will in part be "unnatural". Classic OT, in contrast, is specifically designed to allow only "natural" grammars, constructed by ranking universal and/or functionally motivated constraints.[9] It thus fails to provide an adequate account for how accidents of history are incorporated into synchronic systems. I return to this issue later, but refer the reader to Kiparsky (1973) for detailed discussion of how unnaturalness develops in an RBP grammar.

[9] Several OT supporters now acknowledge the need for parochial/language-specific constraints; cf. Boersma 2000; Ellison 2000; Mohanan 2000; Green 2001, 2005; Hayes and Albright 2003; Bye forthcoming; and in a sense the targeted constraints of Wilson 2000 *et seqq*.

2.1.1.4 *Natural processes: Interlanguage* Interlanguage phenomena that reflect neither the native nor the target language, such as Hungarian- and Farsi-speaking learners of English producing final devoicing (Altenberg and Vago 1983 and Eckman 1984 respectively), do not make sense in OT. Most OT supporters assume that humans start with a default ranking of the universal constraint set, but this can't be what is surfacing in the Hungarian and Farsi speakers' interlanguage, which has neither the ranking of the native language nor of the target language (neither Hungarian, Farsi, nor English has a rule of final devoicing). In OT, once the learner reranks the constraints, the original (=UG) ranking is lost; one therefore predicts the non-existence of interlanguage effects, except for the oft-mentioned emergence of rankings that are underdetermined in the L1, especially in the treatment of loanwords. Since the rankings relevant to coda voicing are determined in Hungarian and Farsi, though, hidden rankings cannot be responsible for the observed interlanguage devoicing. (Uffmann (2004) tries to account for this effect in OT by assuming that second-language learners pass through an initial M >> F stage, but this wrongly predicts that second-language learners should show the same pronunciation patterns as first-language learners.) RBP, on the other hand, allows for second-language learners to postulate rules that are not part of their native or target languages. (This freedom of rule postulation is also essential in explaining spontaneous emergence of crazy rules in first-language acquisition and counterfeeding opacity in second-language acquisition (Idsardi 2002).)

2.1.1.5 *Ineffability* A fifth phenomenon that any theory of phonology must account for is ineffability. Some derivations produce no output whatsoever, such as schm-reduplication with words like *schmo* and *Schmidt* for many English speakers. Two central tenets of OT, Violability and Emergence of the Unmarked, explicitly predict that ineffability should not exist. Orgun and Sprouse (1999) show that the Null Parse account of this phenomenon proposed by Prince and Smolensky (1993) does not work; their own solution however requires abandoning Violability, which seriously undermines the OT enterprise. In RBP, on the other hand, such effects are derived by means of inviolable surface constraints.

Our serial, rule-based model is able to account straightforwardly for each of the five important classes of phenomena outlined above, whereas classic Optimality Theory, wherein as Itō and Mester (1997) put it, "there is no sequential phonological derivation in the sense of traditional generative phonology [and] there is no set of rules and operations applying in a certain order," is fundamentally unable to derive any of them in an insightful way.

2.1.2 *Overgeneration*

The second major problem with OT is that it predicts the existence of unattested phenomena. Steriade 2001 for example observes that some phonological constraints receive only one solution across languages, e.g. devoicing in syllable codas. One of the core tenets of classic OT, free ranking and factorial typology (McCarthy and Prince 1993: 145), explicitly and incorrectly requires that a wide range of repair strategies be employed cross-linguistically to deal with violations of this constraint. I expand on this problem in Sections 2.3 and 2.8.

2.1.3 *Failure to solve RBP problems*

The third major problem with OT is that it fails to provide satisfactory solutions to the problems it identifies in RBP, notably the problem of conspiracies. I return to this issue in Section 2.9.

2.1.4 *Acquisition as generalization formation*

Finally, OT misses the fact that grammar construction is driven by the extraction of generalizations from the data to which the learner is exposed. These generalizations are encoded directly in rules and inviolable constraints, whereas OT is forced to simulate their effects via complicated constraint rankings, which in turn can only be arrived at after comparing the outputs of an equally complicated array of competing rankings. In this sense the learning strategy employed in RBP is formally simpler than what is required in OT, and more insightfully captures our intuitions concerning the nature of the acquisition process.

In the remainder of this chapter I elaborate on the most important of the points outlined above. Before discussing these points, though, I would like to clarify what I am taking as the objects of comparison.

2.2 Definitions

The form of RBP employed here assumes that the surface representation of the morphemes in a sequence is derived from their underlying representations by the application of a series of ordered rules. These rules are subject to the cycle, Structure Preservation, the Derived Environment Condition, and inviolable constraints on underlying and surface representations.[10] The details of this theory are set out in Halle and Vergnaud 1987;

[10] There is a general misconception by OT supporters that RBP does not include constraints of any sort, but inviolable constraints such as the OCP and Final Consonant Extraprosodicity were in

Halle and Marantz 1993; Kenstowicz 1994*b*; Halle 1995; Calabrese 1995; and Vaux 1998.

OT on the other hand is currently a moving target. McCarthy (2000: 149) has stated that "the central thesis of OT is that a grammar is a language-particular ranking of violable, universal faithfulness and markedness constraints." Steriade recently offered a weaker formulation of this, namely that the central element of OT is the idea that constraints can be in conflict, and when they are their outcome is determined by ranking.

To be interesting and falsifiable, though, a theory of grammar must say things (or, more technically, *make predictions*) about human language and human languages. OT as defined by McCarthy and Steriade above says nothing about either of these. Put more starkly, a theory that contains nothing more than the principle of constraint ranking is *uninteresting*; without Richness of the Base, parallelism, factorial typology, and the rest of what is normally called "classic OT," as set out in Kager 1999, OT says nothing about conspiracies, abstractness, etc., because, as McCarthy himself points out, it could be implemented derivationally.

It is only by adding in specific constraints and principles of constraint construction, UR construction, levels (or absence thereof), and so on, that one is able to deal with actual data and thereby evaluate and attempt to falsify the theory. OT supporters therefore tend in practice to employ a more fleshed-out version of OT, which I label "Classic OT," that contains something like the elements in (3).

(3) Classic Optimality Theory (McCarthy and Prince 1993: 144–5; Kager 1999; McCarthy 2002*a*: 109):
universal set of markedness and faithfulness constraints + GEN + EVAL + constraint ranking + strict domination + violability + parameterization via ranking + parallelism of constraint satisfaction + alignment + ROTB

In this chapter the label "OT" generally refers to Classic OT as outlined in (3) and wherever possible to the core set of assumptions common to all forms of the theory; where variation in the theory plays an important role, as in the treatment of opacity, I try to account for the different options.[11]

common use long before the appearance of OT, and continue to be part of most rule-based theories. See Sections 2.7 and 2.9 for further discussion.

[11] Some readers might object that the generalized form of OT evaluated here is not espoused in this particular form by any phonologist, to which I respond that this represents my best attempt to strike a balance between Kager (1999) and the other leading forms of OT, "in an attempt to capture what [is] essential to the [theory], eliminating the inconsistencies and the debilitating unclarities of the various approaches that are developed in the literature. As an interpretation, it might be incorrect;

2.3 Opacity

Thus armed with working definitions of RBP and OT, let us return to comparison of the two theories, starting with the problem of opacity (4). The classic take on opacity (excluding a number of interesting modifications by Baković 2007) comes from Kiparsky (1973: 79):

(4) Opacity according to Kiparsky (1973)

A phonological rule P of the form A → B / C _ D is opaque if there are surface structures with any of the following characteristics:

a. instances of A in the environment C _ D
b. instances of B derived by P that occur in environments other than C _ D

What are the predictions of OT and RBP with respect to opacity? Phonologists generally acknowledge that RBP predicts the existence of opaque rule interactions within grammars cross-linguistically, assuming that children are exposed to data that justify the postulation of opaque orderings. Classic OT on the other hand allows only focus counterfeeding, according to McCarthy (1997*b*). Counterbleeding interactions, which RBP produces in the form outlined in (5), would have to be modeled in classic OT as in (6) (McCarthy 1997*b*).

(5) Counterbleeding

UR	ABC#
B → D / _ C	ADC#
C → E / _ #	ADE#

(6) OT version: *BC >> Faith(B→D); *C# >> Faith(C→E)

/ABC/	*BC	Faith(B→D)	*C#	Faith(C→E)
[ADE] (opaque)		*		*
[ABE] (transparent)				*
[ADC]		*	*	
[ABC]	*		*	

We can see in (6) that the second candidate, [ABE], incurs a subset of the violations of the first candidate, [ADE]. Hence there is no ranking of the as-yet unranked constraints that will yield the first candidate as the output. As McCarthy (1997*b*) points out, classic OT allows only transparent interaction in such cases.

but to reject attempts at such interpretation is pointless, since the only alternative is to reject what exists as inconsistent and vague, overlooking the important insights embedded in it.” (Chomsky 1967: 110)

We have seen so far that RBP predicts the existence of both counterfeeding and counterbleeding opacity, whereas classic OT predicts that only focus counterfeeding should be attested. The actual facts of language support the RBP prediction and not the OT prediction: every known language (as well as many forms of child language and adult interlanguage, as we'll see later) has opacity effects, and the types that Classic OT rules out, including counterbleeding opacity, are in fact quite common. A well-known example occurs in Tiberian Hebrew, where glottal deletion counterbleeds epenthesis (7) (cf. Idsardi 1997).

(7) Tiberian (Masoretic) Hebrew

a. epenthesis into final clusters
/melk/ → [melex] 'king'
/ʔerts/ → [ʔerets] 'land'

b. ʔ-deletion in coda
/qaraʔ/ → [qa:ra:] 'he called'

c. interaction: counterbleeding (Epenthesis >> ʔ-deletion)
/deʃʔ/ → [deʃe]'tender grass' (not *[deʃ])

Two objections that I am aware of have been raised against the RBP treatment of opacity. The first maintains that opaque rule orderings pose a learning problem (cf. Peng 2002). In reality, though, the acquisition scenario for opacity in RBP is simple, as has already been demonstrated formally by Kiparsky (1973) (cf. also Johnson 1984 for formal discussion): the child first learns two independent generalizations, based on an underdetermined data set, and then later, when confronted with data that bring the two generalizations into conflict, makes a decision about how to order them relative to one another. (This process is actually directly analogous to the mechanism by which constraints come to be ranked in OT.) The learning schema just outlined directly produces the range of attested opacity effects. OT, on the other hand, encounters serious learnability problems with respect to opacity, as I discuss at the end of this section.

The second problem claimed for the RBP take on opacity is that it predicts the existence of counterbleeding Duke of York interactions, which putatively do not exist (McCarthy 2003*a*). This turns out not to be a problem for RBP, since several such cases are known to exist; cf. Greek (Newton 1972), Catalan (Bermúdez-Otero 2002), Polish (Rubach 2003), and Karaim (Nevins and Vaux 2004).

Classic OT, on the other hand, by virtue of its monostratal architecture wrongly predicts a large class of opacity effects to be impossible (McCarthy

2002*a*). Given the rampant and undeniable attestation of opacity effects of diverse sorts in the languages of the world, OT supporters have proposed a number of patches, including **local conjunction** (for counterfeeding opacity: Kirchner 1996; Baković 2000; Łubowicz 2002; Moreton and Smolensky 2002; Itō and Mester 2003*a*), **OO constraints** (Benua 1997; Burzio 1998), **sympathy** (McCarthy 1999*b*, 2003*a*), **stratal OT** (Bermúdez-Otero 1999; Kiparsky 2000; Rubach 2000, 2003; Itō and Mester 2003*b*), **turbidity** (Goldrick 2001), **targeted constraints** (Wilson 2001), **comparative markedness** (McCarthy 2003*c*), and **virtual phonology** (Bye 2001). As I outline in the rest of this section, though, none of these patches deals with the opacity problem in a satisfactory manner.[12]

Local constraint conjunction (LCC) makes it possible to derive a subset of opacity effects by teaming a markedness constraint with a faithfulness constraint. By allowing for a potentially unlimited set of constraints to be constructed on a language-specific basis, though, LCC seriously undermines the central OT tenet of Universality, and creates non-trivial learning problems (it is not clear how or when such constraints would be constructed in a learning model such as Tesar and Smolensky's). Moreover, as McCarthy 2003*c* points out, "th[e] greater flexibility of local conjunction is unwarranted and typologically problematic. By conjoining the wrong constraints or conjoining them in the wrong domain, it is possible to produce D[erived] E[nvironment] E[ffect]s [and grandfathering effects] that are not only unattested but quite implausible." Van Oostendorp (2005) elaborates that LCC does not capture the locality of DEEs: a faithfulness violation anywhere in a word combined with a markedness violation elsewhere could generate a DEE, but this is unattested and implausible. McCarthy 2003*c* adds that conjoining the wrong faithfulness constraints can produce impossible results, such as unconditional augmentation and circular chain shifts.

Attempts to deal with opacity via Output-Output (OO) constraints fare no better. McCarthy (1997*b*: 5) points out that such constraints do not work in cases where no form in the paradigm shows the desired phonological process, such as the famous Hebrew form deʃe in (7). Potts and Pullum (2002) add that OO (and sympathy) constraints are not easily expressed using modal logic, and "introduce serious conceptual worries." Kissock, Hale, and Reiss (1998,

[12] I do not consider here McCarthy's (2006) theory of candidate chains (OT-CC) or Pater's (2007*b*) Local Harmonic Serialism, as they essentially concede the opacity problem to the derivational camp. As McCarthy (2006) states in his abstract, "In the revised theory, candidates consist of chains of forms that somewhat resemble the derivations of rule-based phonology." OT-CC moreover predicts the existence of a type of opacity that appears not to be possible in human languages: "OT-CC predicts a type of interaction, referred to as counterfeeding from the past, in which phonological process P1 is able to feed process P2 except when some other process P0 applies earlier in the derivation." (Wilson 2006)

2000) adduce a number of additional problems with OO theory, such as the lack of consistent and explicit principles governing the selection of the base, and making predictions that turn out to be empirically incorrect.

McCarthy's Sympathy Theory is perhaps the most obviously and broadly flawed of the OT attempts to deal with opacity effects. Kiparsky (2000) observes that "once we look at entire phonological systems, not just toy examples of a few interacting constraints, sympathy results in very serious loss of generalization." Idsardi 1997 adds that Sympathy fails to eliminate the existence of conspiracies, the central advantage claimed by OT, adducing examples such as stress shift in Russian and epenthesis and spirantization in Hebrew. (Myers (2002) makes a similar point for the famous *NC̥ constraint, on which see also Blust (2004).) Sympathy moreover creates chaos in systems with multiple opacities (Idsardi 1998; Kiparsky 2000); is unable to deal with opacity of allophonic processes such as nasal harmony in Sea Dayak, rendaku in Japanese, and Canadian French vowel harmony (Poliquin 2006), thanks to the requirements of the rich base and restricting sympathetic constraints to the family of faithfulness constraints (McCarthy 2003*c*, 2005*a*: 28; Itō and Mester 2003*a*: §3.2); relies on otherwise unmotivated constraints and rankings (Kiparsky 2000); predicts non-occurring types of constraint interactions, e.g. mutual non-bleeding (Kiparsky 2000); is unable to distinguish between lexical and postlexical epenthetic vowels (Kiparsky 2000); fails to derive transitivity of opacity (if A is opaque with respect to B and B with respect to C then A is opaque with respect to C; Kiparsky 2000: 14); wrongly predicts that if "two notionally distinct processes … violate exactly the same faithfulness constraints, then they must always act together in rendering a third process opaque" (McCarthy 1999: §3.2; for counterevidence from Hebrew, see Idsardi 1997; Idsardi and Kim 2000; and Levi 2000); is unable to mimic serial derivations requiring two or more intermediate representations, such as the Catalan case discussed by Bermúdez-Otero 2002; fails to capture the link between opacity in non-alternating items (dealt with via sympathy constraints) and paradigmatic misapplication (handled by OO correspondence) (Bermúdez-Otero 2003); is unable to identify a sympathetic candidate in Itelmen epenthesis (Cable 2004); lacks a sensible phylogenetic origin (Bermúdez-Otero 2003), which violates the central OT tenet of functional grounding; violates evaluationism,[13] the essence of constraint-based grammar (List and Harbour 2001); provides no trigger for the acquisition of opaque grammars (Bermúdez-Otero 2003); results in undergeneration (Itō and Mester 1997; de

[13] "Evaluationism [is] the claim that the constraint violation scores of any two candidates contain sufficient information to rank them in a global harmony ordering." (List and Harbour 2001)

Lacy 1998; Bermúdez-Otero 1999, 2003) by confining sympathetic candidates to a subset of those that obey an IO-faithfulness constraint F (McCarthy 1999*b*: 339); requires that the sympathy constraint be invisible to selection of flower candidate (McCarthy 1999*b*: 339; Kager 1999: 391; Bermúdez-Otero 2003); and can depend on cumulativity (McCarthy 1999*b*: §4.2), which is *ad hoc* and wrongly excludes non-paradigmatic non-vacuous Duke of York gambits (Bermúdez-Otero 2003).

Ultimately, Sympathy introduces complexity and disorder without fully addressing the problems it purports to solve. In order to account for German x∼ç allophony, for instance, Itō and Mester (2003*a*) postulate a ranking Max >> *VC that is neither motivated by the transparent phonology nor the default ranking provided by UG, assuming an initial state where M outranks F. "[S]ympathy turns out to be not simply additive to the basic setup of the grammar induced on the basis of the transparent phonology (which surely takes acquisitional precedence). Rather, in order to be workable, sympathy requires further reranking of constraints in order to ensure that basic properties of the language to be generated are still correctly captured" (Itō and Mester 2003*a*, p. 15 in ROA version). In short, "[Sympathy Theory] gets more and more complicated, without succeeding in resolving the existing problems. Old and revised S[ympathy] T[heory] seem to be too perplexing or daunting . . . to be convincing or psychologically plausible." (Coutsougera 2000: 45)

Kiparsky (1997) attempts instead to account for opacity with a particular implementation of Stratal OT that "permits elimination of a type of alignment constraint and of OO, BR, and Sympathy." Although Kiparsky's Stratal OT handles some matters left unsettled by Sympathy (e.g. Japanese rendaku in Itō and Mester 2003*b*), it has its own particular set of problems. Fearing that proposing a multistratal model could be viewed as tantamount to reverting to a derivational system, some phonologists have (based on Koskeniemi 1983) restricted their models to two levels (Orgun 1996*b*; Rubach 2000). Others (notably Goldsmith 1993, Lakoff 1993, and Kiparsky 2000) have included three levels of representation in their frameworks, thereby trivializing the strata, according to McCarthy (1997*b*: 4). Although limiting the number of strata to two avoids a complete relapse into a traditional rule-based system, a two-level approach is not sufficient for resolving the famous Hebrew *deʃe* case (McCarthy 1997*b*). As we saw in Chapter 1, moreover, Orgun (1996*a*) has convincingly demonstrated on the basis of Uyghur data that the phonology needs to be able to have as many cycles as there are affixes, so that one cannot limit the phonology to two or three levels.

If there is no independent motivation within OT for postulating multiple strata, Kager (1999) argues, then introducing strata creates a hybrid framework which would have to be abandoned for the more general derivation model (385). Kager is also puzzled by the lack of dramatically different rankings at different levels (385), given the potential for large-scale rerankings. McCarthy and Prince (1999) agree that "crucial evidence distinguishing serialist from parallelist conceptions is not easy to come by; it is therefore of great interest that reduplication-phonology interactions supply a rich body of evidence in favor of parallelism. Malay ... Southern Paiute ... and other examples cited in McCarthy and Prince 1995 ... either cannot be analyzed serially or can be analyzed only in formally-problematic and conceptually-flawed recastings of conventional serialism" (291). (Raimy 2000 in fact accounts for the phenomena in question within a serialist framework; see ch. 5 of the present volume for further details.)

Setting aside the dilemma of how to incorporate serialism into OT without appealing to a derivational model, we find that there are practical shortcomings in Kiparsky's Stratal OT. Stratal OT allows Duke of York derivations (a problem according to McCarthy (1997*b*: 11) but not in my opinion); results in affix-ordering paradoxes (Sproat 1985, 1988; Halle 1987); and fails to obtain grandfathering effects, derived environment effects, and rules that counterfeed themselves (McCarthy 2003*c*, p. 50 in web version).

In order to account for these latter three phenomena, McCarthy proposes Comparative Markedness (McCarthy 2003*c*), in which a given candidate's markedness is compared with markedness of the most faithful candidate. Unlike stratal OT, Comparative Markedness predicts that all related processes should stand in the same counterfeeding relationship with the process with which they interact (McCarthy 2003*c*, pp. 51–2 in web version). CM also predicts that derived environment effects (which require nM >> F >> oM) and counterfeeding opacity (which require oM >> F >> nM) should not coexist, although there is evidence that they do in Meskwaki (Wier 2004). Like Sympathy and classic OT, CM cannot deal with opacity of allophonic processes, such as nasal harmony in Sea Dayak, because of what ROTB requires (McCarthy 2003*c*); it cannot account for voice inversion in Luo (McCarthy 2003*c*); and it doesn't trigger failure of rendaku in, e.g., sakatoŋe (McCarthy 2005*c*: 28) because of ROTB. OO-CM constraints moreover predict (apparently unattested and implausible) anti-cyclic effects wherein things happen only to forms once they are fully embedded in other forms, without the shape of the outer material being relevant; e.g. Turkish-prime [kitap], [kitep-lar] (van Oostendorp 2003). Finally CM allows circular chain shifts and violations of harmonic ascent in general (McCarthy 2003*c*).

Another attempt to deal with opacity effects within an OT framework is Targeted Constraints Optimality Theory (TCOT) as employed in Wilson 2000, Baković and Wilson 2000, and Chen-Main 2007. Wilson 2000 points out that TCOT and Sympathy Theory deal with opacity in similar ways: a winning candidate is selected on the basis of similarity to an optimal transparent form, modulo a sympathetic/targeted constraint. The key difference between the two treatments of opacity, according to Wilson, is that TCOT, unlike Sympathy Theory, avoids Duke of York effects.[14]

TCOT does in fact contain a kernel of truth, insofar as it introduces into OT analogs of several essential components of RBP (sequential derivations, severely restricted GEN, language-specific constraints and repairs). It also encounters a number of problems, however. Firstly, McCarthy (2002*b*) asserts that targeted constraints do not solve Too Many Solutions problems such as First Consonant Deletion and Coda Devoicing, as Wilson (2001) claims, because the theory of targeted constraints (i) relies on inventory restrictions, which OT does not contain, and can be subverted by inventory-affecting constraints, and (ii) requires possible and impossible UR-SR mappings to incur identical faithfulness marks, which doesn't appear to be possible. Secondly, McCarthy notes that targeted constraints are unable to compare markedness of segments from disparate or epenthetic sources, which markedness constraints must be able to do in order to have the desired range of effects. Thirdly, McCarthy points out that targeted constraints can be trumped by other constraints in cases where the two relevant candidate outputs (one that is favored by a targeted constraint and one that would be expected to be able to win if the relevant constraint were conventional rather than targeted) are equal in faithfulness and equally marked with respect to constraints other than the targeted one (2002*b*: 287; cf. Blumenfeld 2006 for further cases). Fourthly, targeted constraints do not work well in cases where there is more than one attested way of avoiding a given configuration, e.g. a nasal followed by a voiceless obstruent. Myers 2002 (cf. also McCarthy and Pater 2004) observes that "one could posit different targeted constraints for the different ways of avoiding this configuration, but then one would lose the essential insight of Prince and Smolensky 1993 that the avoided configuration is the same in all these cases." Finally, Salting (2005) suggests that vowel height harmony cannot be dealt with by targeted constraints.

Further problems with recent unpublished versions of TCOT are presented by Pater (2003), Pater and McCarthy (2004), and Rubach 2004. Though many of the problems that have been pointed out with TCOT do not involve opacity *per se*, it should be clear from the above discussion that TCOT is unable to deal with OT's too-many-solutions and opacity problems while maintaining the perceived advantages of the OT architecture.

[14] Comparative Markedness (McCarthy 2003*c*) also resembles Sympathy Theory and TCOT insofar as it relies on comparisons to a maximally faithful shadow candidate to mimic certain types of opacity effects.

On top of the problems with individual OT treatments of opacity just described, there are numerous shortcomings shared by all OT treatments. First, Smolensky 1996 (cited by Idsardi 2002) states that special mechanisms like constraint conjunction are not postulated unless warranted by the data. This cannot explain the appearance of counterfeeding and counterbleeding and derived environment effects in second-language acquisition (see Idsardi 2002 for discussion). Second, List and Harbour (2001) point out that "some cases of NonPareto opacity [wherein] the set of violation scores for the optimal candidate of one selection process [is] too similar (in a technical sense) to the set of violation scores for a suboptimal candidate of another selection process [make] it impossible for any aggregation function using only violation scores to determine the right outcome in both cases and thus such cases are not accommodable within any constraint-based grammar."

Third, a recent investigation of the problem of phonological opacity in Optimality Theory, Virtual Phonology (Bye 2001), reveals a novel type of opacity, 'rule sandwiching', which cannot be derived using any of the optimality-theoretic accounts of phonological opacity discussed above. Specifically, three-rule interactions of the form P > Q > R (where > means 'ordered prior to'), where P and Q interact transparently, but R opacifies Q, and P and R introduce identical faithfulness violations, are ruled out by Sympathy Theory. Data from several languages, including Yawelmani Yokuts, Hebrew, Mohawk, and North Saami provide support for the existing of rule-sandwiching effects, contrary to the predictions of OT treatments of opacity.

Fourth, most OT theories of opacity have problems with counterbleeding of the deʃe type. Sympathy and LPM-OT can deal with this class, but each encounters problems of its own, as already discussed. Calabrese (2005) points out that the extra machinery introduced in order to account for opacity brings no additional insights to the theory, whereas RBP accounts for opacity via extrinsic rule ordering, which is independently required in the model. By Occam's Razor, the power of our theory should be extended only if this extension leads to greater insight than is available in the more constrained theory (Calabrese 2005). OT treatments of opacity fail to satisfy this requirement.

Finally, the RBP treatment of opacity is significantly more elegant than its OT counterparts: it predicts exactly the attested types of opacity effects and deals with them straightforwardly and in a unified way (see Idsardi and Kim 2000 for further elaboration and exemplification). Since opacity is one of the most fundamental phenomena in human language, we must prefer a theory that accounts for it straightforwardly (RBP) over one that seems unable to deal with it (OT).

Some supporters of OT have responded that what RBP treats as a unified phenomenon, opacity, is actually a heterogeneous set of unrelated facts that are only made to look like a coherent whole by the theory. My response to this is that, to paraphrase Sampson (1975), one fact needs one explanation. Our linguistic intuition, be we derivationalists or OT supporters, suggests that grammars involve generalizations that may conflict with one another; RBP provides a more successful account for this fact. One could add that, all else being equal, a theory that accounts for a range of phenomena via a single mechanism is to be preferred over a theory that accounts for the same facts with two or more mechanisms.

2.4 Iterativity and cyclic effects

I mentioned earlier that opacity created by iterative rules creates even more profound problems for OT, since proposed patches such as Sympathy, level ordering, and output-output constraints cannot apply (Hyman and VanBik 2002; Wolfe 2000). As McCarthy (2002*a*: 172) states, "within-level opacity, if it exists, will present exactly the same problems for [stratal OT] as it does for classic OT."

The problem for OT is that within-level opacity *does* exist. Consider for example the well-known Abkhaz stress system, outlined in (8)–(11). The basic rule is that Abkhaz assigns word stress to the leftmost (underlying) accented syllable not followed by another accented syllable, and otherwise to the final syllable (Dybo 1977; Wolfe 2000); the effects of this generalization can be seen in (8i–ii).

(8) Abkhaz

a. assigns word stress to the leftmost (underlying) accented syllable not followed by another accented syllable, and otherwise to the final syllable (Dybo 1977; Wolfe 2000).

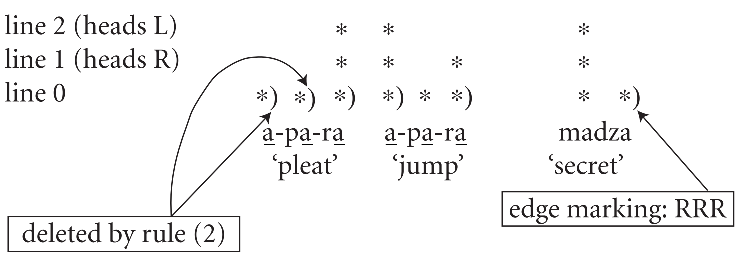
b. lexically accented vowels underlined; surface stresses indicated by an acute accent

i. *verbs*	**accented root**		**unaccented root**	
	$\underline{\text{a}}$-p$\underline{\text{a}}$-r$\underline{\text{á}}$	to pleat	$\underline{\text{á}}$-pa-r$\underline{\text{a}}$	to jump
	$\underline{\text{a}}$-j$\underline{\text{a}}$-r$\underline{\text{á}}$	to lie down	$\underline{\text{á}}$-fa-r$\underline{\text{a}}$	to eat
	$\underline{\text{a}}$-ts$\underline{\text{a}}$-r$\underline{\text{á}}$	to go	$\underline{\text{á}}$-ta-r$\underline{\text{a}}$	to give
ii. *nouns*	madzá	secret	(unaccented root; surfaces with final accent)	
	$\underline{\text{á}}$-madza	def.-secret		
	madzá-k'	secret-indef.		

The fact that the leftmost underlying accent wins results from a familiar iterative rule of Clash Deletion, schematized in (9), which in terms of the Halle-Idsardi stress system deletes the leftmost of a pair of adjacent brackets. Default rightmost stress results from RRR edge marking in tandem with right-headedness on Line 1 of the stress grid.

(9) Clash Deletion:) → Ø / _ *) (Iterative, L → R)

I provide a sample derivatior in (10).

(10)


The key here for our purposes is that Clash Deletion produces edgemost effects, but the domain—an accent sequence—is not a prosodic constituent and therefore is not amenable to interpretation in terms of OT constraints, which we expect by dint of MAX to produce *V̲VV̲ (e.g. á̲-pa̲-ra̲), not VVV̲, from an underlying V̲ V̲ V̲ sequence, as depicted in (11).

(11)

/ a̲-pa̲-ra̲ /	NOCLASH	MAXACCENT	LEFTMOST
* * * * * * ápara	*!*		
* * * * * * ☞ apará		*	*!
* * * * * * 💣 ápara			

Cases of this type cannot be handled in any currently accepted form of OT, because the opaque interactions involved occur within a single level of derivation and therefore cannot be explained away by adding additional levels *à la* Kiparsky (2000), nor can they be attributed to paradigmatic pressures.

It would in theory be possible to generate effects of this type using interlevel constraints of the sort "do not have a bracket in an output form when its correspondent in the input is adjacent to another bracket," but such constraints have been demonstrated by McCarthy (1997*b*) to create significant problems for OT in other areas, and therefore should be excluded from the universal constraint set. (See also Kager (1999) for discussion of how two-level constraints do not work for Oromo compensatory lengthening.)

In sum, RBP again handles this sort of opacity straightforwardly, though this time the formal device involved is simple iterativity. In OT, on the other hand, this sort of intralevel opacity poses a serious problem.

2.5 Optionality

Now let us turn to optionality. Like opacity, optionality is not predicted by the architecture of Classic OT. Numerous devices have been proposed to deal with this problem within an OT setting, including but not limited to underdetermination (Hammond 1994), cophonologies, tied constraints (Anttila 1997*b*), and differential constraints (Horwood 2000), but all of these fail to account for the entire range of optionality effects. Most notable of these is sequential optionality (also called "local optionality" (Riggle and Wilson 2006)), which results from the interaction of optionality with iterativity.

To see how this works, let us return to the topic of predictions. The form of RBP endorsed here allows rules to be marked as [±optional] *and* as [±iterative]. This being the case, we predict that it should be possible for a rule to be marked as *both* [+optional] *and* [+iterative]. Such a rule would produce a nuanced type of optionality wherein both options for a rule, application and non-application, can appear within a single word.

On the other hand, in classic OT, which does not contain the [±optional] and [±iterative] variables, we predict only all-or-nothing optionality: a process should either apply or not in all of the environments in which its structural description is met. This is precisely what we find with Warao labial voicing (Howard 1972: 87): /p/ optionally surfaces as [b], but if it does then *all* ps in the word must surface as [b], as shown in (12). (RBP also predicts the existence of all-or-nothing effects: these result from optional rules that are [-iterative].)

(12) Sequential optionality: labial voicing in Warao (Howard 1972: 87)

- /p/ optionally surfaces as [b]

- If it does, then all ps in the word must surface as [b]
- [papa] ∼ [baba] (*[paba], *[bapa])

It is important to note that the all-or-nothing effect is actually a direct consequence of one of the most central components of OT, parallelism, according to which entire, fully formed outputs are evaluated in parallel. In this system it is not possible for the constraints to peek at the intermediate workings of GEN, and hence heterogeneous outputs are emphatically predicted to be impossible.

In actual fact, though, heterogeneous outputs of the sort predicted to be impossible by OT *do* exist, just as we expect in RBP. Though statistical frequency is irrelevant for our purposes, such processes are actually quite common. One such example is English flapping (13). Though the precise environments for the flapping rule are difficult to pin down, its application appears to be optional when the target is flanked by two unstressed vowels, as in the word *marketability*. In RBP, the flapping rule is marked as both [+iterative] and [+optional], and therefore applies in the following manner. It proceeds directionally through the word, say left to right, scanning for an alveolar stop that meets its structural description. When it finds one, in this case the t at the end of *market*, it then either applies or not, depending on the outcome of the algorithm responsible for optionality. It then moves on to the next potential target, in this case the t of -ity, and again either applies or doesn't. Crucially, though, the choice of whether or not to apply to the second t *is independent of* the choice that was made for the first t. This is a necessary consequence of the rule being iterative, and actually appears to make the correct empirical prediction for the flapping rule, as can be seen in (13).

(13) Optional allophony in free variation: predictions for *marketability* with regard to English flapping

a. RBP: [mɑɹkətʰəbɪlətʰi] ∼ [mɑɹkəɾəbɪləɾi]∼[mɑɹkətʰəbɪləɾi]∼ [mɑɹkəɾəbɪlətʰi]

b. OT: [mɑɹkətʰəbɪlətʰi]∼ [mɑɹkəɾəbɪləɾi]

c. Actual outputs (for my idiolect): [mɑɹkətʰəbɪlətʰi]∼ [mɑɹkəɾəbɪləɾi]∼ [mɑɹkətʰəbɪləɾi]∼ [mɑɹkəɾəbɪlətʰi]

Similar results hold for English glottalization in words like *continental*.

In their discussion of optional complementizers in English, Baković and Keer (2001) argue that optionality originates from the richness of the base. The multiple surface forms are derived from multiple input forms and not from the application of optional constraints or rerankings on a single underlying form. When faithfulness contraints outrank markedness constraints, these multiple input forms yield multiple outputs. In their analysis, ROTB allows all four forms listed in (13a) to be inputs, each of which then surfaces as an output.

The most famous case of sequential iterative optionality involves the French rule of schwa deletion, as discussed in a series of publications by François Dell. Dell shows that this rule optionally deletes schwa following a VC sequence, proceeding iteratively from left to right within a phonological phrase and subject to familiar phonotactic restrictions. Just like in the English flapping case, the combination of iterativity and optionality in French schwa deletion produces heterogeneous outputs; since French allows long strings of schwas, though, the heterogeneity is even more striking than in English, as shown in (14b), where a single string of four schwas produces a set of eight outputs.

(14) French schwa deletion

a. ə→ Ø / V (#) C _, L→R , optional across #

b. envie de te le demander 'feel like asking you' (Dell 1980: 225)

ãvidtəldəmãde
ãvidtələdəmãde
ãvidtələdmãde
ãvidətələdmãde
ãvidətlədmãde
ãvidətlədəmãde
ãvidətəldəmãde
ãvidətələdəmãde

A curious variation on the theme of sequential optionality appears in Dominican Spanish as described by Núñez Cedeño (1988) (see also Bradley 2006). This dialect possesses a rule that optionally inserts /s/ at the end of a syllable; the rule applies iteratively, which again produces a range of outputs for a given polysyllabic input. This rule differs from the English and French equivalents, however, in applying only once per word; in other words, it proceeds iteratively through a word looking for a target and then applies optionally to that target, but it appears that once the rule actually applies to one of its targets it then stops. The effects of this rule can be seen in (15b).

(15) Optional s-epenthesis in Dominican Spanish (Núñez Cedeño 1988)

a. Ø → s / _ $]_\sigma$ (optional, structure-preserving)
b. /abogado/ 'lawyer' → asbogado, abosgado, abogasdo, abogados

It should be clear that *none* of the three types of sequential optionality just discussed can be accounted for in Classic OT, which can produce only all-or-nothing effects. Donca Steriade (pers. comm.) has suggested that the French facts might be obtainable if one assumes variable construction of prosodic phrases, but we have no independent evidence for this variation,[15] nor will this trick work in the English case, where the prosodic conditions for the two targets are identical, or in the Dominican Spanish case, where phrasing does not appear to be involved.

One might also try marking constraints as optional, implementing this by having constraint evaluation proceed iteratively through a word, with EVAL then having for each target the option of assigning an asterisk. It is not clear that the Dominican facts can be derived in this way, though, and this strategy moreover undermines the spirit of the OT enterprise (Cheryl Zoll, pers. comm.).

Riggle and Wilson (2006) propose Local Optionality to account for some of the examples described here. Unlike Global Optionality, in which a process applies across the board in a single form, Local Optionality allows the existence of position-specific constraints. In addition, the reranking of these contraints occurs within rather than across derivations. Local Optionality, though able to account for a number of examples of optionality, including schwa deletion in French, fails to adequately explain the all-or-nothing effects we find in phenomena such as Warao labial voicing.

Boersma and Hayes (2001) propose to derive optionality via interaction between overlapping constraints. In their Gradual Learning Algorithm, constraints are associated with a range of values on a continuous ranking scale. The ranges of two constraints can overlap, leading to variation in the ranking of those constraints at the time of evaluating a particular UR/candidate set mapping, resulting in some cases in the selection of more than one output for a given input. Because these ranges are implemented as probability distributions, the free variation produced by the algorithm will be similar to the free variation found in the training data. Boersma and Hayes offer their algorithm as an alternative to Tesar and Smolensky's (1996, 1998) Constraint Demotion, which many (including Tesar and Smolensky themselves) have observed cannot effectively deal with optionality (Boersma and Hayes 2001). Boersma and Hayes's model has its own shortcomings as well, though, such as not producing local optionality effects, predicting unattested variation in metathesis in Ilokano glottal stop deletion (Horwood 2000), and being unable to converge on the correct analysis in cascading credit problems (Pater 2005).

2.6 Exceptionality and unnatural processes

Consider next the problem of naturalness. Classic OT inherits from Natural Phonology the belief that synchronic phonological systems are "natural," i.e.

[15] Until we have evidence that French phrasing is more complex than Dell assumes, the RBP analysis is to be preferred on grounds of parsimony.

everything in them makes synchronic sense. I argue to the contrary that phonological grammars can be "unnatural," as noted by Bach and Harms 1972, Kiparsky 1973, Anderson 1981, Hayes 1996, Blevins 1997, McMahon 1998, Hyman 2000, Calabrese 2005, and others. In fact, as Anderson points out, careful scrutiny reveals that *most* of the phonology of natural languages is non-natural. As I suggested in 2001, unnatural systems of this type are accounted for most efficiently and insightfully in a Chomskyan rule-driven framework. Existing OT implementations can be altered to account for the relevant phenomena, but only at the cost of abandoning the central theoretical tenets that have been claimed to give them the advantage over derivational theories. This loss of insight is inevitable, since OT is specifically designed to account for the (supposed) fact that all phonology is natural. Put in general terms, the search for explanation in language will not find everything in synchronic structure, just as natural selection does not explain everything in nature. In both areas, much of the explanation is to be found in history, as was already noted by Chomsky 1966 and Kiparsky 1973.

To give this debate substance, let us consider the example of productive phonological consonant epenthesis, which is frequently maintained by OT supporters (Lombardi 1997, Steriade 2001, etc.) to employ only default consonants like homorganic glides, ʔ or h. Contrary to this belief, the Turkic language Uyghur employs consonant epenthesis in several situations, including the ones in (16a) and (16b), but regardless of the quality of neighboring segments chooses y or r, rather than any of the natural choices prescribed by OT.

(16) Uyghur (Hahn 1991: 25)

a. y inserted between two vowels at morpheme boundary
oqu+Al- → [oquyal-] 'to be able to read'
iʃlä+Al- → [iʃläyäl-] 'to be able to work'

b. y inserted between CV root and C suffix
yu:-b → yuyup 'wash and . . . '
su:-m → süyüm 'my liquid'

The seemingly unexpected selection of [y] and [r] as epenthetic segments can be directly connected to the fact that precisely these two segments undergo optional deletion in syllable codas (Hahn 1992: 77, 79), as can be seen in (17).

(17) optional r- and y-deletion in syllable coda

a. kördüm ∼ ködüm 'I saw'
bazar ∼ baza 'bazaar'

b. päyzi ~ päzi 'gorgeous'
hoyla ~ hola 'courtyard'
eytiŋ ~ etiŋ 'tell!'

Most interesting for our purposes is the possessive paradigm in (18), where we find alternation between y and r as the epenthetic consonant in cases that do not involve underlying /r/ or /y/.

(18) Uyghur possessive marking (Hahn 1992: 90)

	a. stem	**b. 1st person /-m/**	**c. 2nd person /-ŋ/**	**d. 3rd person /-(s)i/**
girls	qɪz-lar	qizlirim	qizliriŋ	qizliri
mother	ʔana	ʔanam	ʔanaŋ	ʔanisi
spring	baha:(r)	baharim	bahariŋ	bahari
street	kotʃa	kotŋam	kotʃaŋ	kotʃisi
ink	siya:	siyayim/siyarim	siyayiŋ/siyariŋ	siyasi
chicken	toxu:	toxuyum/toxurum	toxuyuŋ/toxuruŋ	toxusi

We can see in columns b and c of (18) that the first-person suffix /-m/ and the second-person singular suffix /ŋ/ attach directly to stems ending in short vowels, but give rise to a [+high] epenthetic vowel when following consonant-final stems. The third-person singular suffix in column d is also underlyingly consonant-initial, but in postconsonantal position this /s/ deletes rather than triggering epenthesis, as with [qizliri] rather than *[qizlirisi].

The interesting property of Uyghur for our purposes is that it avoids superheavy syllables. Adding monoconsonantal suffixes such as -m and -ŋ to stems ending in long vowels such as toxu: 'chicken' should produce forms containing superheavy syllables such as *toxu:m, but outputs of this type are ungrammatical. Uyghur chooses instead to epenthesize twice, yielding forms such as toxuyum and toxurum in (19); according to Hahn [y] and [r] are in free variation in these situations.

(19) /toxu:-m/ → [toxuyum] ~ [toxurum], not *[toxu:m]

Why are [y] and [r] chosen for insertion here rather than say glottal stop or a homorganic glide? The variation between [y] and [r], which are precisely the segments that delete in the complementary environment in Uyghur, clearly demonstrates that these two segments are chosen for insertion because they are also targets of deletion. No manipulation of the feature specifications of y and r in tandem with homorganic glide insertion can save the day here.

One might try instead to say that all long-vowel roots have been historically reanalyzed as ending in y or r, and it is this y or r that surfaces in columns b and c in (18). This analysis runs into a number of problems. First, it requires systematic hypercorrection of all long-vowel roots, with subsequent postulation of y- and r-final allomorphs for every single long-vowel stem, which relegates to the domain of arbitrary lexical content something that otherwise receives a simple phonological explanation. Second, as Hahn (1992: 90) observes, "If an inserted y or r had become an underlying segment in a given root, then such a root would be expected to take on the allomorph –i" in the third person, which it does not, as shown by minimal pairs like *bahari* vs *siyasi*. Finally, notice that forms with underlying /r/ such as *bahar* do *not* show the y∼r alternation, but instead surface with an [r] in all situations where it is not placed in a syllable coda during the course of the derivation: *baharim*, *yarim* 'my dear', etc.

It is therefore clear that Uyghur employs both r and y insertion, and that the choice of these particular segments results not from homorganic glide insertion but from hypercorrection. In other words, a synchronically arbitrary segment is chosen for insertion by a completely regular phonological rule for reasons that are ultimately historical. RBP correctly predicts that language learners will be able to postulate unnatural rules of this sort if exposed to the right kind of evidence, whereas universalist implementations of OT wrongly predict that they should be unlearnable.

In addition to the problem with unnatural rules, Calabrese (2005) notes that we also require idiosyncratic language-specific negative constraints in order to account for accidental gaps, such as the absence of the unmarked vowel /u/ in Huave (Noyer 1994) or of non-palatalized č in Russian.

2.7 Ineffability

Our next major phenomenon that runs counter to the predictions of OT is absolute ungrammaticality, or what is sometimes called "ineffability." One of the most robust cross-linguistic generalizations is that some derivations produce no output whatsoever; for example, many speakers of English find no output of schm-reduplication to be grammatical with schm-initial words like schmo and Schmidt (Nevins and Vaux 2003). In Vaux and Nevins's (2007) online survey of schm-reduplication, 128 out of 300 (43%) respondents preferred a null output for schmuck, 117/300 (39%) for schmooze, and 126/300 (42%) for Schmidt.[16]

Phenomena like this are easily analyzed in RBP, which has at its disposal inviolable output constraints with the power to crash a derivation. Within OT an analogous move is more troublesome, because it violates the central tenet of OT that *all* constraints are violable (Prince and Smolensky 2002: 6).

[16] Many other respondents opted for avoidance strategies not easily accounted for without postulating allomorphy: 17 respondents selected shluck/shlooze, 4 selected fluck/flooze, and 1 selected vluck/vlooze.

In order to deal with this problem, Prince and Smolensky (1993) propose that Gen produces a special candidate called the Null Parse, which lacks a morphological category and has no phonetic realization. This candidate is stipulated ("ex hypothesi" in the words of McCarthy and Wolf 2005) to satisfy all well-formedness and faithfulness constraints and to be the only candidate output that violates the special constraint MParse, which requires that all underlying forms have a surface realization.

Prince and Smolensky designed these propositions to ensure that any constraint C ranked above MParse would in effect be inviolable, because any candidate that violated C would lose to the Null Parse, as shown in (20).

(20) candidates:
X violates constraint C but not MParse
Ø the Null Parse candidate: violates MParse but not C

/input/	C	MParse
X	*!	
Ø		*

When the Null Parse ends up being selected as the optimal output the surface result is ineffability, as the ill-formedness of the Null Parse makes it "uniquely unsuited to life in the outside world" (Prince and Smolensky 1993: 51) and unpronounceable.

The Null Parse analysis encounters several serious problems. First, the stipulation that the Null Parse candidate satisfies all well-formedness and faithfulness constraints appears to be arbitrary and unmotivated by independent principles (Nevins and Vaux 2003; Rice 2005; see McCarthy and Wolf 2005 for an attempt to make these stipulations follow from revisions to Correspondence Theory). Second, the Null Parse analysis fails to capture the intuition that an output with no phonetic realization is qualitatively different than no output at all. In the case of *schm*-reduplication, for example, our intuition is that reduplication produces the output *schmuck-schmuck*, and this output is then discarded because it violates a constraint requiring that the base and the reduplicant be distinct. Our intuition crucially does *not* suggest that *schm*-reduplication produces an output with no phonetic content, which therefore is not pronounced. (Orgun and Sprouse 1999 make the same point with regard to Swedish *[rätt].) McCarthy and Wolf (2005) respond, building on Coetzee's (2004) theory that the output of Eval is not a single optimal candidate but rather a ranking of all candidates for relative harmony, that what is being accessed in conscious assessments of ineffability phenomena is not a rejected winner but rather a first runner-up to a victorious but ineffable Null Parse candidate. Resorting to this level of counter-intuitive, theory-internal sophistry when a simpler theory directly produces the attested facts and intuitions strikes me as one of the clearest indications that OT has gone down the wrong path in our quest to understand phonological cognition.

Rice (2007) points out that the expected form is in fact being produced, but only surfaces when the right phonological conditions are present. For instance, Norwegian imperatives generally consist of the bare verb root (21a). When the root ends in a consonant clusters of rising sonority, though, this form is not allowed to surface, resulting in a null output (21b).

(21) Norwegian imperatives
  a. /spis/ 'eat' → spis 'eat!'
  b. /padl/ 'paddle' → *padl 'paddle!'

Related to this is the fact that negative imperatives generally allow the negative *ikke* to surface on either side of the verb (22a), but only postverbally with rising-sonority roots, where the initial vowel of *ikke* is able to license the final consonant of the verb root as its onset (22b).

(22) Norwegian negative imperatives
  a. hopp ikke på møblene ~ ikke hopp på møblene 'don't jump on the furniture!'
  b. klatr ikke på møblene ~ *ikke klatr på møblene 'don't climb on the furniture!'

Similarly, rising-sonority infinitives are allowed when immediately followed by vowel-initial but not consonant-initial prepositions: *sykl opp bakken* 'bike up the hill!' but **sykl ned bakken* 'bike down the hill!'.

A third problem with the Null Parse analysis involves ranking paradoxes, as demonstrated for Turkish by Orgun and Sprouse (1999). In Turkish, suffixed forms must contain at least two syllables (23); ungrammatical monosyllabic forms are not augmented by epenthesis as one might otherwise expect (24).

(23) Suffixed forms in Turkish must contain at least two syllables

root	**gloss**	**suffixed form**	**gloss**
sol	the musical note G	sol-üm	my G
do:	the musical note C	*do:-m	my C

(24) Ungrammatical monosyllabic forms are not augmented by epenthesis
*do-yu-m (cf. /araba-a/ 'car-dat' → [arabaya])
*do:-u-m (cf. /el-m/ 'hand-my' → [elim])

The ranking required to generate these effects is Dep, Lex=PrWd, FtBin ≫ MParse. There is a problem with this ranking, though: the ranking Dep ≫ MParse predicts that epenthesis should never be possible in Turkish, when in fact it is possible (cf. *arabaya* 'to the car', *elim* 'my hand'). We thus have a ranking paradox: the null output for 'my C' (23) requires Dep ≫ MParse, but the epenthesis in *arabaya*, *elim*, etc. requires MParse ≫ Dep.

Orgun and Sprouse remedy this problem by suggesting that individual constraints may be specified as inviolable, "Control" constraints in a given language (cf. also Pesetsky 1997, 1998; Fanselow and Féry 2002). As McCarthy (2005*b*) rightly points out, though, their solution requires abandoning Violability, which seriously undermines the OT enterprise and the notion that EVAL is a total function (one that yields an output no matter what the input is). McCarthy therefore attempts to explain away the facts adduced by Orgun and Sprouse, but the other problems just discussed remain.

For example, it should be clear from the English schm-reduplication and Norwegian infinitive cases that the expected winner is running afoul of an inviolable constraint, something that is easily capturable in RBP and in OT endowed with Control. Rice (2007) demonstrates moreover that one can produce paradigmatic gaps using Optimal Paradigms theory (OP; McCarthy 2005*b*). Invoking OP raises a number of new problems, however.

Firstly, OP compounds the already serious extension problem raised by the GEN+EVAL component of OT. It seems unlikely from a computational and psychological perspective that speakers generate and consider vast numbers of possible outputs each time they produce a word, especially when compared to a theory (RBP) that accounts for the facts equally well or better and does not encounter the extension problem. OP compounds this problem by requiring that the selection of the surface form of a word involve generation and evaluation not only vast numbers of candidate outputs for the underlying representation of that word, but also of every permutation of the set of those outputs and the candidate outputs for other members of the paradigms to which that word belongs.[17]

[17] Kautz and Selman (1991) show that the problem of determining whether a given default non-monotonic theory has an extension is highly intractable (NP-complete, to be precise), seemingly

Secondly, one of the two central predictions of OP is falsified by Trukese and Yiddish data. McCarthy (2005*b*) points out that OP predicts the impossibility of true underapplication within paradigms, because "OP has the same basic logic as base-reduplicant identity, so it similarly predicts that underapplication is only possible in inflectional paradigms when overapplication is ruled out by some high-ranking constraint." Cable (2004: 17) shows that underapplication of Trukese minimal word-induced vowel lengthening cannot be attributed to a higher-ranking constraint blocking the relevant overapplication candidate, thereby falsifying this prediction of OP. Albright (2004) shows that the underapplication prediction is falsified by Yiddish loss of final devoicing as well. (He deals with the underapplication effect in Yiddish by having bases in inflectional paradigms, which loses some of the claimed advantages of McCarthy's model.)

Thirdly, Bobaljik (2006) demonstrates that it is morphosyntactic category and not paradigm properties that determine phonological behavior in cases of the sort discussed by McCarthy (2005*b*).

Fourthly, Rice (2005) observes that "the motivation to have fewer violations [in OP] effectively rewards paradigms with gaps ... Taking this line of reasoning to its absurd extreme, the evaluation of paradigms by constraints referring to the markedness or faithfulness of phonological properties of the members of the paradigms will reward the paradigm with the most gaps. Indeed, a paradigm with gaps in every cell—the null paradigm—will be optimal." Rice remedies this with Max{CAT} constraints requiring realization of morphological categories, but his revised version of OP encounters problems as well.

Specifically, Rice (2007) points out that his analysis predicts that "if there is a phonotactic problem in two different potential words (infinitive, imperative, etc.) within the same category (verb, noun, etc.), they must be repaired in the same way." A case that may fit the description of what Rice predicts to be impossible involves the manifestation of root-initial geminates in Homshetsma, as described by Berens (1997). Homshetsma, a variety of Armenian spoken in northeastern Turkey, contains a verb /t^h:-/ 'hit', whose geminate /t^h:/ surfaces as such only in intervocalic position within a prosodic phrase (25a-c). When a preceding vowel is not available to license the first half of the geminate, the gemination surfaces on an immediately following consonant if one is available (25d). Elsewhere, i.e. if no consonantal host

because the problem requires checking all possible sequences of firings of defaults (cf. Antonelli (2006) and discussion in Section 2.10 of this chapter).

is available for the gemination (Homshetsma has no long vowels), the /t^{h}:/ surfaces as a singleton (25e).

(25) Manifestations of underlying geminates in Homshetsma

	UR	gloss	SR
a.	/gu-t^{h}:-a-m/	'I hit (pres.)' (imperfective-hit-theme.V-1sg)	[gut^{h}:om]
b.	/mi-t^{h}:-a-Ø/	'don't hit' (prohibitive-hit-theme.V-2sg.neg)	[mit^{h}:a]
c.	/indzi t^{h}:-a- Ø mi/	'don't hit me' (me hit-theme.V-2sg.neg prohib.)	[indzi t^{h}:a mi]
d.	/t^{h}:-v-i/	'I hit (past)' (hit-passive-1sg.aorist)	[t^{h}ev:i]
e.	/t^{h}:-u/	'hit!' (hit-2sg)	[t^{h}u]

The Homshetsma data are thus a problem for Rice because, contrary to the explicit prediction of his model, a single phonotactic problem (syllable-initial geminates) triggers three different repairs within the same grammatical category.

Putting together this problem with what we have seen in the rest of this section, no known version of OT is able to deal with the robust empirical problem of ineffability without (in the case of Control Theory) abandoning one of the central tenets of the model, Violability, or (in the case of the Null Parse analysis, Rice's version of OP, and McCarthy and Wolf's 2005 theory) creating a swathe of incorrect predictions and psychological and computational problems. In contrast, the form of RBP assumed in this paper encounters no such problems, by virtue of containing inviolable constraints (cf. n.9).

2.8 OT predicts the existence of unattested phenomena

Let us now move on from phenomena that OT predicts not to exist (or at least cannot derive in a straightforward or insightful way) to the converse, phenomena that OT predicts to exist but do not. Given free ranking (e.g. McCarthy and Prince 1993: 145, Kager 1999; Section 1.7; McCarthy 2002*a*: 109; Féry and Fanselow 2002: ch. 3) and the absence of an appropriately restrictive general theory and inventory of constraints, it is strange that many OT papers start from the observation that a given phenomenon doesn't exist. To the contrary, without a constrained set of constraints almost anything can exist.

But what exactly is the overgeneration problem for OT? On top of the problematic predictions identified by Colin Wilson that we discussed in Section 2.3, there is the specific subtype of overgeneration that McCarthy terms "overkill" (also known as "the too many repairs problem" or just "overgeneration"), wherein a constraint violation is repaired cross-linguistically in only a subset of the ways one might expect. In light of recent OT interest in this issue, it is ironic that OT supporters initially attacked RBP for linking target and repair and thereby constraining—purportedly without reason—the set of possible repairs (cf. (10)). In recent years, though, the OT supporters have "discovered" that the set of repairs is in fact constrained in certain ways, as has been maintained all along in RBP.

Consider the example of vowel deletion discussed in Casali (1997). He claims (p. 509) that at "the boundary between two lexical words ... the constraints violated by V1 elision constitute a subset of those violated by V2 elision. In these contexts, therefore, I predict that only V1 elision is possible." His claim is empirically incorrect, as shown by languages such as Sanskrit, where word-initial *a*- deletes after word-final mid vowels (Whitney 1889: 47). Even if such languages did *not* exist, the fact remains that it would be easy to come up with an OT constraint system that would generate exactly the behavior that Casali claims to be impossible. In this case OT is *right* to allow for more possibilities than some of its proponents are aware of, but there is a more general problem, the "too many solutions problem": OT allows for a wide range of grammars that appear to be impossible (Steriade 2001; Lombardi 2001). Let us consider two examples, one from Steriade and one from Flemming.

Steriade (2001) observes that some phonological constraints receive only one solution across languages; for instance, she claims that the constraint punishing [voice] specifications in codas is invariably dealt with by devoicing. This generalization is incorrect—see Eckman (1981) and Edge (1991) for counterevidence and Vaux (2005*b*) for discussion—but let us imagine for the sake of argument that it is correct. Free ranking explicitly and incorrectly predicts that a wide range of strategies should be employed cross-linguistically to repair violations of this constraint.

Flemming (2001) observes along similar lines that "not all conceivable rankings of MAXIMIZE CONTRASTS correspond to possible languages. The balance between maximization of the number of contrasts and maximization of the distinctiveness of contrasts is determined by the ranking of MAXIMIZE CONTRASTS relative to the MINDIST constraints. Allowing all definable rankings predicts the existence of languages which value the number of contrasts very highly, resulting in a huge number of very fine contrasts, and languages which value distinctiveness very highly, resulting in a handful of maximally distinct contrasts. Neither of these extremes is attested."

He continues that "It seems that there is a lower bound on the distinctiveness required for a contrast to be functional, and that there is an upper bound beyond which additional distinctiveness provides a poor return on the effort expended. This could be implemented by specifying that certain MINDIST constraints, referring to the smallest acceptable contrastive differences, are universally ranked above MAXIMIZE CONTRASTS, and that MAXIMIZE CONTRASTS is in turn universally ranked above another set of MINDIST constraints which make 'excessive' distinctiveness requirements. However it would be desirable to derive these bounds from general considerations of perceptibility and communicative efficiency rather than simply stipulating them."

How should one deal with overkill problems of these types? The combination of unfettered GEN and free ranking make this a non-trivial problem in OT. Steriade (2001) proposes to constrain certain types of repair via the P-Map, a matrix of confusion-based similarity indices, but this solution is unsatisfactory for many reasons and is unlikely to cover all cases of overgeneration, particularly those that do not involve perceptual similarity.[18] In RBP, on the other hand, a relatively simple solution is available, since one is able to limit the inventory of repair strategies provided by UG. A plausible theory of this sort is developed in Calabrese (2005).

A further fertile avenue for constraining repairs (and typology in general) has been investigated in numerous articles by John Ohala (most of which are available at http://trill.berkeley.edu/users/ohala/index3.html), Hale and Reiss (2000), Blevins (2004), and Vaux and Samuels (2004, 2005). The basic idea of what I call the Ohala Theory is that many or all of the patterns we find in phonological systems are actually products of history—itself the product primarily of phonetic constraints and influences on the acquisition process—and need not, and in fact *should* not, be assumed to constitute part of synchronic grammars.

Steriade and Baković (in personal communications) have rightly pointed out that the Ohala Theory is not the exclusive property of RBP; it can be incorporated just as well within an OT framework. Though Steriade and Baković are technically correct, by saying they could use the Ohala Theory—but don't—they are using what Postal (2004: 292) terms the "Psychic Alternation Move": "this criticism of A's claim is not valid, because although A admittedly made the claim, he could easily have made a different, correct claim instead."

Why do OT supporters resort to the Psychic Alternation Move instead of adopting the Ohala Theory? I believe the primary reason is that the spirit

[18] See Blumenfeld (2006) for further discussion of problems with P-Map theory.

of OT for most OT supporters (and inherited, I would argue, from Natural Phonology) involves incorporating function, origins, and explanation in general into the synchronic grammar, which is not compatible with the Ohala Theory.

In this context it is important to bear in mind when considering using negative typological evidence the following quote:

> Certain apparent linguistic universals may be the result merely of historical accident. For example, if only inhabitants of Tasmania survive a future war, it might be a property of all then existing languages that pitch is not used to differentiate lexical items. Accidental universals of this sort are of no importance for general linguistics, which attempts rather to characterize the range of possible human languages. The significant linguistic universals are those that must be assumed to be available to the child learning language as an a priori, innate endowment.
>
> (Chomsky and Halle 1968)

Applying this notion to our present topic, I would suggest that the overkill problem is not a problem at all; consequently, OT supporters (and in fact all phonologists) should cease building theories (such as Steriade's P-Map and Wilson's TCOT) on (often faulty) negative typological evidence.

There is, however, one respect in which overkill may pose a legitimate problem for OT: to the extent that the theory by virtue of combining free ranking and a virtually unbounded inventory of constraints allows for virtually limitless possible systems, it is conceivable that a subset of these are ones that are not actually considered by children acquiring a language. This sort of mismatch, in the spirit of the Chomsky and Halle quote above, is one we should actually try to exclude.

2.9 OT fails to provide satisfactory solutions to the problems it identifies in RBP

Now that I have addressed the major problems concerning what OT predicts to be possible and impossible, I would like to address the fact that OT fails to provide satisfactory solutions to the problems it identifies in RBP, most notably conspiracies (cf. (1g)). As Kisseberth (1970) first observed, the basic problem with conspiracies is that the application or non-application of multiple phonological processes sometimes appears to be guided by a unitary output goal. Kisseberth then suggests that "by factoring out the target from the individual rules ... we convert the generalization inherent in the conspiracy into a formal simplification. Given that formal simplicity is taken as the basis of the evaluation measure, we thereby succeed in characterizing grammars as more highly valued insofar as they have conspiracies." (Kiparsky 1973: 59)

Calabrese (2005) points out that the ability to provide a single formal device to generate a conspiracy, namely a constraint, hardly constitutes an advantage for OT over RBP. It is true that the form of RBP addressed by Kisseberth in 1970, namely that of SPE, did not employ inviolable constraints in a prominent fashion. By 1993, though, most rule-based theories employed a suite of inviolable output constraints, such as the OCP, which were perfectly capable of generating conspiratorial effects. It is therefore unclear why OT supporters identify conspiracies as a problem for RBP.

Idsardi (1998, 2000) observes moreover that OT itself is still forced to postulate conspiratorial analyses for phenomena such as English r-deletion, stress shift in Russian, and Hebrew epenthesis and spirantization. I would add to this that single constraints never account for conspiracies on their own; one always needs at least two constraints operating in tandem to produce a given conspiratorial output (cf. McCarthy's (1998) recasting of opaque rule orderings in terms of ranked constraint pairs, discussed in Chapter 1). In the famous Yawelmani case, for example, production of a surface light syllable from an underlying cluster requires collaboration between not only the NoCoda markedness constraint that is commonly implied to underlie the conspiracy, but also a specific pair of Max and Dep faithfulness constraints, all three of which must be ranked in a specific manner with respect to one another in order to simulate the effects of the relevant rules.

We must also be careful to avoid reifying what superficially look like they may be language-internal conspiracies of the Yawelmani variety but are actually composites of independent phenomena in separate languages. As Blust (2004) points out in the context of his critique of Pater's (1999, 2001) work on *NC:[19]

The notion of a conspiracy appears to be defensible so long as the evidence supporting it is drawn from a single language ... One and the same conspiracy can, of course, be found in different languages, but if the argument for functionally related processes consists entirely of *comparative* data, the nature of the argument is fundamentally changed. Some 6,000 languages are still spoken, many with significant dialect differences, and the number of sound changes or synchronic residues of sound change is therefore at least 6,000 times what one can expect to find in a single language. Given the range of choices, it is hardly surprising that in different languages or language

[19] Blust points out literally dozens of further empirical and conceptual problems with Pater's conspiracy analysis of the Austronesian data; I refer the reader to Blust's original article for detailed discussion.

families one can find completely unrelated phonological processes that happen to overlap in eliminating a particular input. Yet this is precisely the form that many arguments in OT have taken in recent years, and for this reason they are precariously speculative.

I would also like to suggest, following Kiparsky (1972), that we should not be so quick to assume that a given set of processes is controlled by a conspiratorial global rule or constraint. Kiparsky (1972 (=1982*a*: 112)) suggests instead that elements putatively implicated in conspiracies, such as "three-consonant clusters, adjacent stresses, and so on, are linguistically complex configurations, and rules eliminating or avoiding them are accordingly highly natural and occur frequently in the languages of the world. It is therefore only to be expected that there should be some languages in which several rules should eliminate or avoid these configurations, and that there should be languages in which no instances of these configurations appear on the surface . . . What I am questioning, then, is whether there is any fundamental sort of difference between the cases in which just one or two rules reflect general phonological conditions of this type, and the cases in which several rules are involved, which would be termed a 'conspiracy'." He then adds, "concrete empirical differences are clearly also involved: for example, is there any evidence for a true 'functional unity' of the rules in a conspiracy which would not simply be characterizable by their sharing a common target? Are there cases in which they are subject to parallel historical changes at some point in the development of a language? Are there cases in which apparently diverse changes in the rules of a language at some point in time can be shown to be consequences of the imposition of a single derivational constraint? Are there cases where the rules in a conspiracy have the same set of lexical exceptions? This would be strong evidence in favor of derivational constraints. However, I have not found any such cases."

Kiparsky outlines several further formal objections to the conspiracy theory that remain relevant today. First, there is the problem of indirect participation in a conspiracy—cases where a rule participates in a conspiracy indirectly, by appropriately feeding or bleeding another rule. Kiparsky states that in order to deal with such cases "we would therefore have to say something like this: a rule [sc. constraint—BV] is highly valued (or 'free') if its application creates representations to which other rules [sc. constraints] are applicable in such a way as to implement the conspiracy." Second, Kiparsky points out (1982: 114) that the formal devices by which an output constraint can be effected are highly heterogeneous. Therefore, factoring out those parts of the structural analyses of processes involved in the conspiracy is technically feasible only in a small part of the relevant cases. Third, Kiparsky mentions ordering paradoxes from Kenstowicz and Kisseberth 1970 as phenomena that might require derivational constraints, but then states that they might have functional underpinnings (1982: 114–15).

Given these problems with Conspiracy Theory, Kiparsky then develops (1973) a sort of selective evolutionary account for the appearance of supposed conspiracies, in which opaque rule systems are less likely to be acquired successfully by language learners, and conspiratorial rule orderings, being relatively transparent, are therefore more likely to survive than their opaque competitors. This line of thinking is quite compatible with the mechanisms expounded for the emergence of phonological typologies in Evolutionary Phonology (2004) and appears to be on the right track.

Closely related to the conspiracy problem is the so-called Duplication Problem, which refers to the isomorphism between Morpheme Structure Constraints and phonological rules that is sometimes called for in derivational analyses. Here Anderson (1974: 292) provides a similar explanation to Kiparsky's, which also finds echoes in the more recent work of Ohala, Hale, and Reiss: "the reason a language contains both a morpheme structure constraint of a given type and a phonological rule which results in much the same constraint applying to derived structures, though the two are distinct, is that both serve to enforce some natural constraint. Both the constraint and the rule, that is, have the same explanation, where an explanation in phonological terms is often provided by our substantive empirical knowledge of the physics and physiology (and perhaps, eventually, neurology) of speech." He adds that "both the constraint and the rule require independent statement in the grammar, since each may have (independent) idiosyncracies," a line of reasoning also raised by Kiparsky. Anderson concludes that "as far as the formal apparatus of a description is concerned, then, we see no alternative to positing separate rules and conditions of morpheme structure. The attempt to unify a rule and a constraint (or two rules) is not, properly speaking, a job for phonological descriptions."

There is another sort of Duplication Problem that does *not* arise in RBP but *does* afflict OT. Mohanan (2000) observes that if two processes within or across languages differ just in the domain of application of a pattern, such as nasal assimilation within vs. across words in Malayalam, OT is forced to split the pattern into two distinct constraints so that the two parts can be ranked differently. This requires an unnecessary and unwanted duplication of the same constraint. If one considers more than two parallel cases the situation becomes even worse; Mohanan demonstrates for instance that in order to account for place assimilation in English, Hindi, and Malayalam, OT would be forced to split place assimilation into five distinct universal constraints. Crucially, each of the five constraints triggers the exact same process; an important linguistic generalization is therefore being missed. Mohanan observes that it was precisely this sort of duplication of a single generalization that led Halle (1959) to reject the classical phonemic level of representation; the same logic should apply to the OT case.

To sum up this final section, I have suggested that OT attributes to RBP problems that are not actually problems, that OT itself fails to solve these problems, that the problems themselves may not exist, and their apparent effects have plausible historical and physiological explanations that do not require duplication in the grammar.

2.10 Conclusions

To conclude this chapter, I have argued that OT has failed to surmount the problems its practitioners associate with Rule-Based Phonology, and also creates new insurmountable problems. The adoption of OT leads moreover to serious loss of generalization in many core areas.

When to OT's problems of undergeneration (e.g. sequential iterative optionality, ineffability, and crazy rules), overgeneration (e.g. Wilson's nasal blocking of epenthesis and counterfeeding from the past), and loss of generalization (especially concerning opacity) one adds the problems of unconstrainedness (Calabrese 2005), unrealistic modeling of linguistic performance (Clements 2000), indeterminacy (Clements 2000), substance abuse (Hale and Reiss 2000), and constraint duplication (Mohanan 2000), and the fact that the self-proclaimed OT successes in accounting for markedness, naturalness, and conspiracies are not an exclusive OT prerogative—see for example the derivational theory developed in Calabrese (1995, 2005)—one sees no reason to maintain OT in face of a descriptively and formally superior rule-based model.

To this conclusion one might add the larger computational problem that OT is non-monotonic (Besnard, Fanselow, and Schaub 2003), by virtue of the fact that one can override conclusions by adding new premises (constraints and rankings).[20] OT thus stands in opposition to generative theories such as RBP insofar as the point of a generative grammar (qua formal computational system) is precisely that it is decidable whether a string is well-formed or

[20] The rule for monotonicity is that if $\Gamma \vdash \varphi$ and $\Gamma \subseteq \Delta$ then $\Delta \vdash \varphi$ (if φ is a consequence of a set of premises Γ, then it is also a consequence of any set Δ containing Γ as a subset). Non-monotonic logics are logics for defeasible reasoning, as monotony is what in classical logic bars one from overriding conclusions by adding new premises; the non-monotonicity of OT is thus closely intertwined with the system of violable, ranked constraints.

not. Though defeasible logics such as the one OT assumes appear to have appealing applications to certain types of real-world scenarios (cf. Dresher 1996), they pose serious and thus far unsolved computational problems such as being NP-complete (Kautz and Selman 1991; Idsardi 2006) and requiring incomputable consistency checks (Antonelli 2006). As Antonelli (2006) states, "Non-monotonic logics appear to be stubbornly intractable with respect to the corresponding problem for classical logic."

RBP, on the other hand, is eminently and efficiently computable. It has been known at least since Johnson (1972) that the effects of phonological rewrite rules of the sort employed in RBP can be simulated using relatively straightforward finite-state machinery, with iterative application accomplished by sending the output from one transducer to the input of the next.

Given the empirical, formal, and computational superiority of RBP *vis-à-vis* OT that we have seen in this chapter, it should now be clear that the parallel constraint-based architecture that currently dominates phonological theory should be abandoned in favor of a serial rule-based architecture. But what then do we do about Orgun's (1993) assertion that serial derivations are cognitively implausible (cf. (1L))? Putting aside for the moment the fact that McCarthy (2006) and Pater (2007*b*) have introduced serial derivations in OT, I respond to the serialism objection that, as Calabrese (2005) observes, human behavior is set in a temporal continuum and therefore requires the acquisition and implementation of ordered sets of instructions. There is no reason for excluding knowledge of serial ordering of instructions, which is fundamental to so many human skills, from the realm of phonology. Itō and Mester (2003*b*, p. 20 in web version) already acknowledge the need for staged strata of phonological computation: "the monostratalism of strict parallel versions of OT undeniably has restrictiveness in its favor, as far as weak generative power is concerned. But the simultaneous loss of descriptive and explanatory adequacy is too high." The facts and arguments adduced in this chapter suggest that we must push phonological theory even further; as Clements 2000 puts it, "many areas of higher-level cognition are admittedly sequential in nature, and it may simply be the case that phonological competence is one of these."

3

Ordering[1]

DAVID ODDEN

3.1 Background

One of the most obvious differences between standard derivational phonology and Optimality Theory is that, at least based on an initial impression, OT should not be able to capture certain relations between inputs and outputs which could be captured in derivational phonology, because OT is not assumed in standard conceptions of the theory to have "intermediate stages" or derivations. This chapter explores some of the machinery available to OT to account for data which is explained by rule ordering and similar derivational devices. In the OT literature, this issue has been addressed under the rubric "opacity." Devices which have been used within OT to replace serial derivations include the Parse/Fill approach, two-level constraints, output-output constraints, and Sympathy Theory, not to mention limited derivationality in the form of level ordering. The chapter focuses particularly on two-level constraints, Sympathy Theory, constraint conjunction, and the use of abstract domains.

To put into perspective the question of how OT should handle facts classically handled by serial derivation, McCarthy (1997*b*) shows that the earlier parse/fill approach and output-output constraints do not give a general account of rule ordering in OT, and says with respect to sympathy theory (p. 18) "Arguably, this is *all* that is required to analyse observed opaque interactions." Kiparsky (1999) appears to make a similar appeal to the desire for a unified account of opacity when he notes that Base-Output constraints cannot explain the opacity of interaction between stress and epenthesis in Bedouin Arabic, where a final syllable underlyingly super-closed by a cluster

[1] This chapter is based on material from papers presented at GLOW 1998, the MIT-Harvard Phonology 2000 Workshop, the Montreal-Ottawa-Toronto Phonology Workshop, and at the Universities of Leiden, Trondheim, and Tromsø, written during 1999–2000 under the auspices of a Fulbright Fellowship at the University of Tromsø.

of consonants receives stress, even though the syllable is opened on the surface by vowel epenthesis.

(1)	/al-walad/	→			[ál-walad]	'the boy'
	/al-himl/	→	al-híml	→	[al-hímil]	'the load'

As Kiparsky points out, this variety of opacity cannot be handled by Base-Output constraints, since there is no surface form lacking the epenthetic vowel to serve as the foundation for anomalous stress in [alhímil] (although it happens that Sympathy constraints can be called on in this case).

I will show here that, well-intentioned desires to limit the number of ways of handling rule ordering notwithstanding, quite a number of devices actually play a crucial role in handling serial phenomena. No unified treatment of rule ordering is possible in OT, and instead one must use a range of devices to accommodate "opacity." This is unsurprising: were there to be a unified account of rule ordering in OT, one might suspect that the device being used is notational trickery allowing one to translate one concept into another, as one can translate upper-case letters into lower-case letters or vice versa. The question I address is whether it is possible to account for all derivational phenomena, once we have marshaled the relevant descriptive machinery, or will it turn out that there are cases which completely resist a non-derivational analysis. There does turn out to be a way in OT to handle virtually all of the cases that I discuss here, though there is a theoretical price to be paid in many of these cases, since a relaxing of theoretical strictures is required. Whether or not this amounts to an unconscionable opening of the theoretical floodgates, or is simply a minor but necessary adjustment within the theory, can only be decided conclusively after a prolonged investigation of a range of derivational phenomena, which is beyond the scope of this chapter. For instance, in order to account for the pattern of H deletion in Kerewe, a two-level account, with two cycles through Gen and Eval, is necessary. Positing a limited derivational aspect to OT does increase the power of OT, but it is unclear whether it amounts to surrendering any fundamental principles of the theory (since it is unclear what principles in OT are truly fundamental versus convenient assumptions). My primary aim is to point out some of the relevant cases, and consider what it takes to handle them.

3.2 Disjoint predictions

A rather basic question about rule ordering which has received little attention in discussions of OT is, simply, just what is the benefit of eliminating derivational steps? One might imagine that some kind of simplification of grammars could result, by eliminating the possibility of extrinsic rule ordering. Allowing rules to be explicitly ordered, under the standard theory of linear rule ordering, allows a given set of *n* rules to be mapped onto *n*! grammars (thus a set of 5 rules maps onto 120 grammars and 8 rules maps onto 40,320 grammars).[2] The consideration of reducing the number of possible grammars is fully negated by the fact that there is an equivalent complication in grammars in the form of constraints ranking, and insofar as a dozen or so constraint may be required to express what a single rule expresses, an OT grammar may actually fare much worse in terms of the combinatorics of basic elements defining grammars, since it is certain that a complete OT grammar requires orders of magnitude more constraints than there are rules in a rule-based account.[3] Another possible motivation for getting rid of rule ordering and serial derivation would be based on the assumption that serial derivations might have a kind of expressive power which is not actually needed to describe natural languages. Creating a tighter fit between theoretical prediction and actual languages is an admirable goal, but it is far from clear that there is anything that couldn't be handled by OT, once all of the necessary machinery is identified.[4]

Another question to be considered is whether there are things that could be accounted for in OT which could not be accounted for in standard derivational phonology, and which do not exist in languages—the implicit assumption has been that, lacking rule ordering, OT is somehow less powerful than a derivational account. I consider two such cases here, one involving Sympathy Theory and the other involving standard devices of OT. In the first example, involving Sympathy Theory, there are rule interactions (transitivity violations)

[2] In OT, constraints are not strictly ranked and are only partially ordered. With partial ordering (ranking), 5 constraints can be ranked 4,231 ways and 8 constraints can be ranked 431,723,379 ways. The general function for computing the number of partial orderings is not known, and values are only known up to $n = 14$ (98,484,324,257,128,207,032,183).

[3] In its own right, it is not particularly important whether an OT grammar requires an order of magnitude more constraints than rules, since there is no theory-independent way of judging the absolute complexity of an analysis. There is at present no reasonable basis for estimating the approximate ratio of rules to constraints, since complete OT-based analyses of languages are rare, and even in extended descriptions of languages in OT, many crucial constraints are implicit in the analysis (e.g. metathesis is a very efficient way to eliminate phonotactically bad sequences, but is rarely used, and yet very few analyses explicitly rule out metathetic candidates). A rule-based account, if technically correct, is necessarily complete up to the limits of the descriptive domain circumscribed by the account. Neither OT nor derivational theory posit intrinsic limits on the number of rules or constraints that may define a grammar, short of the obvious facts that the set of rules/constraints must be finite, and that constraints/rules are not posited without reason. However, a consequence of having more constraints is that there are more ways to order them and thus more possible grammars.

[4] For instance, McCarthy (1997*b*, 1999*c*) argues that OT is incapable of expressing "Duke-of-York" derivations, but it is shown in Section 3.4 that (non-trivial) DY derivations are well within the reach of OT.

that are impossible under standard derivational phonology; however these data can easily be accounted for under Sympathy Theory.

Under standard linear ordering, if rule A precedes B and B precedes C, then A precedes C (and therefore C does not precede A). Commonsensical as this may be, this is an empirically testable hypothesis, and it is possible to construct hypothetical languages which require such a prohibited interaction. The imaginary language Kalaba cannot be modeled in derivational theory, since it involves exactly this contradiction. First, the language has a rule (2), deleting glottal stop before another consonant, which we see applying before the plural suffix *pa*.

(2) *Kalaba Glottal Deletion* ʔ→∅ / ___ C

lim	'tongue'	lim-pa	'tongues'
to	'child'	to-pa	'children'
laʔ	'fish'	la-pa	'fishes'

Second, there is a syncope rule shown in (3) deleting high vowels in a doubly open syllable.

(3) *Syncope* V → ∅ / VC ___ CV
[+hi]

wali	'rice'	wal-pa	'rices'
kapu	'basket'	kap-pa	'baskets'
ugli	'porridge'	ugli-pa	'porridges'
tungu	'onion'	tungu-pa	'baskets'
neno	'word'	neno-pa	'words'

Finally, in (4) there is regressive voicing assimilation that affects obstruents.

(4) *Voicing assimilation* C [-son] → [αvoice] / ___ C [αvoice]

nam	'meat'	nam-gu	'my meat'
nam	'meat'	nam-pa	'meats'
naz	'coconut'	nas-pa	'coconuts'
zig	'load'	zik-pa	'loads'
tot	'infant'	tod-gu	'my infant'

The pairwise ordering of these rules is shown in (5). First, Glottal Deletion must precede Syncope. Glottal Deletion in the second form on the right makes the first syllable open, which then allows Syncope to apply to the vowel *i* in 'fishes.' Second, Syncope must feed into Voicing Assimilation, as shown by the second set of data, where deletion of the high vowel allows *p* to assimilate to *g* in the first example, and allows *z* to assimilate to *p* in the second example.

Thus we deduce the orderings 'A precedes B' and 'B precedes C,' which in standard derivational theory entails that A precedes C.

(5) Glottal Deletion Before Syncope (A > B)

to	'child'	to-pi	'little child'	to-p-pa	'little children'	← /topipa/
laʔ	'fish'	la-pi	'little fish'	la-p-pa	'little fishes'	← /laʔpipa/

Syncope Before Voicing Assimilation (B > C)

kapu	'basket'	kab-gu	'my basket'	←/kapugu/
dizi	'banana'	dis-pa	'bananas'	←/dizipa/

The third set of examples in (6) show that, nevertheless, Voicing Assimilation precedes and is counterfed by Glottal Deletion, that is, C must precede A. These examples show that Voicing Assimilation does not take place between obstruents if they were underlyingly separated by a glottal stop, meaning that glottal deletion has not taken place at the stage where Voicing Assimilation applies.

(6) Voicing Assimilation Before Glottal Deletion (C > A)

to	'child'	to-ʔ	'old child'	to-gu	'my old child'	← /to-ʔ-gu/
tot	'infant'	tot-ʔ	'old infant'	tot-gu	'my old infant'	← /tot-ʔ-gu/
naz	'coconut'	naz-ʔ	'old coconut'	naz-pa	'old coconuts'	← /naz-ʔ-pa/

Thus, the non-language Kalaba is predicted to be impossible in standard derivational theory. The data in (7) shows that Syncope could not be cyclic, which is the one circumstance that might allow for such a rule interaction.

(7)

dizi	'banana'	nu-dzi	'the banana'
mu-n-dizi	'in the banana'	ni-m-nu-dzi	'it is in the banana'

Following standard assumptions about the relation between morphological structure and cyclicity, the following would be the cyclic bracketings and outputs.

(8)

[dizi]	[nu [dizi]]	[mu [nu [dizi]]]	[ni [mu [nu [dizi]]]]
↓	↓	↓	↓
dizi	nudzi	*munudzi	nimnudzi

To derive the correct pattern, Syncope must apply at the word level, iterating from left to right.

While Kalaba is an impossible language in derivational theory, Sympathy Theory allows this language to be described rather easily. The core constraints driving deletion and assimilation are *HeteroVoice, Syncope, and *ʔC (it does not matter whether these are single constraints or sets of constraints which achieve a particular result). The tableaus in (9) show how the non-problematic interactions can be handled trivially, since ordering Glottal Deletion before Syncope, and Syncope before Voicing Assimilation corresponds to transparent satisfaction of all constraints.

(9) *HeteroVoice: *[αvoice] [-αvoice]
Syncope: *VCVCV
*ʔC

	laʔpipa	*HetVoi	Sync	*ʔC
	lapipa		*!	
	laʔpipa			*!
☞	lappa			

	kapugu	*HetVoi	Sync	*ʔC
	kapugu		*!	
	kapgu	*!		
☞	kabgu			

The Sympathy account in (10) provides us with an easy way to describe the opacity of the interaction between Glottal Deletion and Voicing Assimilation, which is simply that voicing has to be faithful to the flower candidate which loses no consonants. Since the candidate preserving all underlying consonants also has a voiceless consonant, the winning candidate must respect the voicing value of that candidate.

(10)

	tot-ʔ-gu	Ident-voi$_{Max\text{-}C}$	*HeteroVoi	*ʔC	Max-C
❀$_{Max\text{-}C}$	totʔgu		*	*!	
☞	totgu		*		*
	totgu	*!			*

Thus Sympathy Theory allows for interactions which cannot be modeled under standard derivational theory. Taking into consideration the further conditions on Sympathy Theory proposed in McCarthy 1999*d*, note that the flower candidate is identical to the input, and thus it has a null set of unfaithfulness mappings in the sense defined in that paper. The tableau in (11) shows that the new interpretation of the assessment of violation of sympathy constraints changes nothing, and the form that is impossible for derivational theory to get is still possible under Sympathy.

(11)

	tot-ʔ-gu	❀Sym	*HeteroVoi	*ʔC	❋Max-C
❀Max-C	totʔgu		*	*!	
☞	totgu	$Max_ʔ$	*		*
	todgu	$Max_ʔ$! $Ident_d$			*

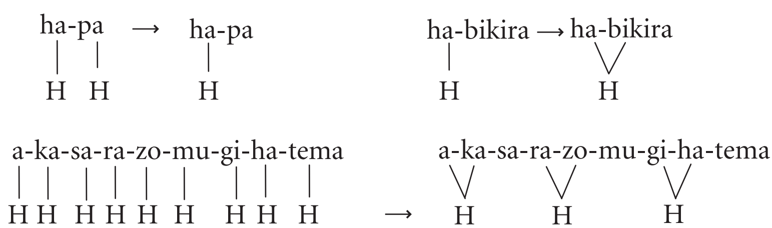
Robust examples of such ordering relations have not emerged from forty years of derivational phonological research,[5] a lacuna which is an underappreciated problem in Sympathy Theory, and more generally with discussions of how OT can get various rule-ordering cases. That is, certain aspects of the OT account of derivationalism result in unjustified generative power, reducing attractiveness that there might have been in a strictly parallel model.

Even without the power of Sympathy Theory, it is easy to show that OT predicts hypothetical process interactions which could not arise under standard derivational theory, and, importantly, which do not arise in natural languages. A number of Bantu languages have a dissimilative tone deletion called Meeussen's Rule which deletes an H after an H, and this process will be discussed for Kikerewe in the next section. Deletion of H after H is a consequence of the OCP, where deletion is an active repair for OCP violations. Another common tonal process in Bantu is rightward Tone Doubling, where H spreads once to the right, eliminating singly linked H tones. In a number of languages with Tone Doubling, a following H tone blocks the rule which is an effect of the OCP as well. (12) illustrates an interaction between these processes with hypothetical data from the imaginary language Kintupú. We will further

[5] Anderson 1969 *inter alia* proposes an alternative theory to rule ordering: local ordering, which allows violations of the transitivity assumption, based on data from Old Icelandic, Faroese, Kasem, Sundanese, and Sanskrit. Most of these cases have been addressed and shown not to argue for 'local ordering', cf. Vroman 1972 for Old Icelandic, Phelps 1973 for Kasem and Sundanese, Phelps and Brame 1974 for Sanskrit. I am not aware of an explicit account of Faroese, and will not essay a full reanalysis of the data provided by Anderson; I will outline the bare essentials of the claim, and my objection to that analysis. The argument for non-transitive ordering in Faroese is based on the claim that a rule hardening geminate glides applies to the output of a rule inserting a glide after a prevocalic high glide or vowel, but only if the triggering vowel is not made to be prevocalic due a rule of intervocalic spirant deletion: hence, /bú-a/ ⟶ *bíwwa* ⟶ [bɪgva] 'to dwell', but /týða/ ⟶ [tuja] 'translate'. The proposed analysis depends on abstract underlying representations with abstract segments /ð/ and /ɣ/, and a rather specific acount of the glide / obstruent alternation exemplified by *bɪgva* 'to dwell' ~ *bujr* 'he dwells'. A crucial assumption of the argument is that this alternation involves hardening of a geminate glide, but no evidence is given for a geminate glide, and indeed Anderson does not consider an alternative analysis where the underlying form contains /gv/ which undergoes preconsonantal vocalization.

assume that Tone Doubling does not spread H to a prepausal syllable, a very common restriction on this process; this restriction allows us to determine that the second of two Hs does indeed delete, as in the first example. The second example illustrates spread of H rightward by one syllable. The third example illustrates the interaction between these processes.

(12)



What should be noticed in the mapping from input to output is that in the third example, a sequence of Hs, two out of every three H tones end up being deleted. This pattern can be described easily in OT. The crucial constraints are the constraint against HH, and a constraint against monosyllabic H domains. In the imaginary language Kintupú, these two constraints are undominated, and the tableau in (13) shows how the correct form is selected, by satisfying both of these constraints at the expense of Max-H. To guarantee that tones delete rather than fusing, Uniformity must also dominate Max-H.

(13)

a-ka-sa-ra-zo-mu-gi-ha-tema H_1H_2 H_3H_4 H_5 H_6H_7 H_8 H_9	*HH	*Mono-H	Unif	Max-H
a-ka-sa-ra-zo-mu-gi-ha-tema | | | | | | | | \ / $H_1H_2H_3$ H_4H_5 H_6 H_7H_8 H_9	*!*******	********		
a-ka-sa-ra-zo-mu-gi-ha-tema \ / \ / \ / \ / \ / H_1 H_3 H_5 H_7 H_9	*!***			****
a-ka-sa-ra-zo-mu-gi-ha-tema $H_{1,2,3,4,5,6,7,8,9}$			*!********	
a-ka-sa-ra-zo-mu-gi-ha-tema \ / H_1				******!**
☞ a-ka-sa-ra-zo-mu-gi-ha-tema \ / \ / \ / H_1 H_4 H_7				******

A derivational analysis of such processes would be founded on two rules, Meeussen's Rule which deletes H after H, and a rightward Tone-Doubling rule, which is blocked from spreading H to a syllable before an H.

(14) *Meeussen's Rule*: H → Ø / H __

Tone Doubling:

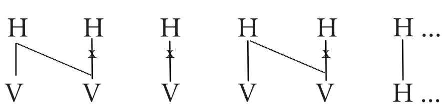
(blocked by H on following syllable; target nonfinal)

The possible outputs from these rules are specified in (15), given either right-to-left or left-to-right iteration in each rule, and either of the possible rule orderings.

(15)	a.	MR(r-to-l) TD(r-to-l)	(ákasarazomugihatema) [ákásarazomugihatema]	→
	b.	MR(r-to-l) TD(l-to-r)	ákasarazomugihatema [ákásárázómúgíhátémá]	→
	c.	MR(l-to-r) TD(r-to-l,l-to-r)	ákasárazómugíhatéma [ákasárazómugíhatémá]	→
	d.	TD(r-to-l,l-to-r) MR(r-to-l)	ákásárázómúgíhátémá [ákasarazomugihatema]	
	e.	TD(r-to-l,l-to-r) MR(l-to-r)	ákásárázómúgíhátémá [ákasárazómugíhatémá]	→ = c.

The pattern of retaining one tone and deleting two following tones, as was easily described under OT, ends up not being describable with ordered rules. Nor could one construct some new rule to perform this operation in one step, along the lines of (16).

(16)



This 'rule' has numerous properties which are prohibited by the general theory of rule construction. First, the rules have to refer to structurally non-adjacent elements. Second, the rule must simultaneously affect multiple foci (in principle, an unbounded sequence). Third, this rule is not even a well-formed rule, insofar as the expression '. . . ' has no formal status in the theory. The theory of rule ordering and rule formulation makes specific restrictive predictions about the interaction of processes, predictions not shared by OT. Lacking any indication that such processes are actually found in human language, this constitutes excessive power on the part of OT.

3.3 OT machinery for reconstructing rule ordering

I now turn to considering some of the formal machinery that will be needed to replace derivational concepts within OT.

3.3.1 *Two-level constraints*

One of the earlier devices proposed to replace derivations is two-level constraints. Two-level constraints were originally proposed by Koskeniemi (1983), and applied by Lakoff (1993) and Karttunen (1993). The essence of a two-level rule is that it refers simultaneously to the input and output stages, so that for example an input element X is mapped to an output element Y just in case it is preceded by Z in the input. Two-level constraints have been proposed in OT, for example in McCarthy (1996) and Orgun (1996*c*). While some cases of two-level constraints such as Bedouin Arabic have succumbed to reanalysis in terms of Sympathy Theory, there remain cases where Sympathy Theory just does not have the necessary power, and therefore Sympathy Theory cannot be the general theory which accounts for canonically derivational concepts. In the previous section, we have considered the tonal dissimilation known in the study of Bantu languages as Meeussen's Rule (MR), where an H tone is deleted after another H. In nearly all Bantu languages with this rule, such as Kikerewe (spoken in Tanzania), every H except the first in a sequence of underlying H tones gets deleted. Languages which exhibit this pattern of tone dissimilation include Kikerewe, Jita, Tonga, Rimi, Kihunde, Nilamba, Luganda, and Haya. An example of this process from Kikerewe is seen in (17). Here, each of the prefixes /táá/, /tú/, /gí/, /kú/ and the first syllable of the stem /hééleezye/ are underlyingly H toned (see Odden (2000) for details of the Kikerewe tonal system). On the surface, each of those H tones except the leftmost is deleted, and the surviving H spreads once to a following non-final syllable by a general tone-doubling process.

(17) /abatáá-tú-gí-kú-hééleezye/ → abatáá-tu-gi-ku-heeleezye
'they who didn't give it to us for you (remote)'
(surface [abatáátúgikuheeleezye] because of low-level tone spreading)

In a derivational account in (18), this pattern is regulated by deleting tones from right to left, working through the string of Hs.

(18) abatáá-tú-gí-<u>kú-héé</u>leezye → abatáá-tú-<u>gí-kú</u>-heeleezye →
abatáá-<u>tú-gí</u>-ku-heeleezye → aba<u>táá-tú</u>-gi-ku-heeleezye →
abatáá-tu-gi-ku-heeleezye ([abatáátúgikuheeleezye] by rightward spreading)

The problem in the OT account in (19) is that way more Hs are deleted than are minimally required to avoid adjacent Hs. The prediction of the OT approach is that only every *other* H tone should delete, since that is the least radical way to eliminate OCP violations. Minimal deletion of H would incorrectly result in the alternating pattern of H tones found in the second candidate.

(19)

	abatáátúgíkúhééleezye	*H H	Max-H
	abatáátúgíkúhééleezye	HHHH	
💣	abatáátugíkuhééleezye		HH
☜	abatáátugikuheeleezye		HH!HH

Nothing useful is added by an appeal to Sympathy Theory, there being no obvious sympathy-inducing constraint. However, as seen in (20), a two-level approach to the constraint on Hs proves to be quite useful. Rather than just prohibiting the appearance of a surface H after a surface H, we can instead prohibit a surface H which stands after an *underlying* H. By stating the constraint this way, deleting every other H is an ineffective strategy for avoiding violation of the constraint, since only the first surface tone in an underlying string of Hs would not end up violating the constraint, given that the *surface* value of the first tone is not considered in computing whether the constraint is satisfied.

(20) */H/ H 'Surface H may not be preceded by an underlying H'

	abatáátúgíkúhééleezye	*/H/ H	Max-H
	abatáátúgíkúhééleezye	HH!HH	
	abatáátugíkuhééleezye	HH!	HH
☞	abatáátugikuheeleezye		HHHH

What this shows, then, is that both two-level constraints and Sympathy Theory are going to be needed in OT (on the presumption that Sympathy Theory is independently motivated).

This two-level version of MR actually helps OT to handle other rule-ordering effects in the language, since MR is not a surface-true principle of the language. In a derivational account, MR is ordered before processes which create HH sequences that do not undergo MR. For instance, an underlying toneless vowel may intervene between two Hs, and by processes of syllable fusion or tone shift, the Hs can become adjacent. In such a circumstance there is no deletion of the second H. As indicated in (21a), the persistive tense is formed with the toneless prefix -*a*- and the H toned prefix *kí*-, which undergo syllabic fusion. Although fusion brings together two H tones on the surface, the second H is not deleted. With the right wording of the constraint, failure of the second H to delete can be accounted for in an OT account, on the basis of the two-level nature of the motivating constraint, since the surface HH sequence does not violate the two-level version of MR. The second H is underlyingly preceded by an underlyingly toneless TBU, /a/, and while that TBU has a surface H, it is only the underlying representation that is considered in determining constraint violation with respect to the first tone in the sequence. Similar non-deletion is illustrated in (21b), involving resyllabifications resulting from optional deletion of an intervocalic glide.[6]

(21)

a. /a-ki-a-kalaanga/ → [achááákálááanga] 'he is still frying'

H H → H H

b. tibákú!yásáma ~ tibákwáásáma 'they aren't opening the mouth'

/ti-ba-ku-yasama/ → ti-ba-ku-asama → [tibakwaasama]

H H → H H → H H

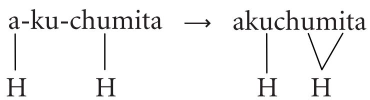
[6] The location of association within the syllable between tone and moras of a long vowel (written as double vowels) is non-contrastive, and the simplest account of this is that the syllable is the structural bearer of tone in the language, hence if any mora in the syllable has H tone, the entire syllable has H, and vice versa. Underlyingly, adjacent vowels in different morphemes which are later syllabified together can individually bear H, as is the case of (21). Utterance-penult long vowels can have a falling pitch which is notated here as *vv*, but this is due to principles of phonetic implementation which will be investigated in a separate study.

Finally, onsetless syllables cannot bear H tone in the language,[7] and underlyingly H-toned syllables undergo rightward tone shift—see Odden (1995). The initial subject prefix of verbs in relative clause tenses is underlyingly H toned (e.g. /bá/), but if the prefix is onsetless (e.g. /á/), then the H shifts to the following syllable (and spreads to the right).

(22)	/bá-ku-baziila/→ [bá-kú-baziila]	‘they who are sewing’
	/á-ku-baziila/→ [akúbáziila]	‘he who is sewing’

Note in (23) that an H tone which is thus shifted to stand before another H does not trigger MR.

(23)	/á-ku-chúmita/→ [a-kú-chúmíta]	‘he who is stabbing’
	/á-ka-kálaanga/→ [a-kú-káláanga]	‘he who is frying’



In all of these cases, there is no deletion of H after H since the two Hs are not underlyingly adjacent, and thus the surface HH sequence does not violate MR. Thus the mechanism that was required to handle the non-minimal pattern of H deletion also accounts for this set of what in a derivational theory would be counterfeeding interactions.

Derived HH sequences are not uniformly exempt from MR. One source of derived HH sequences that are subject to MR involves the interaction between MR and a phrasal tone insertion process, Lapse Avoidance, which inserts H at the end of a noun which stands before a toneless modifier within the phrase, illustrated in (24). This inserted H then spreads rightward because of Tone Doubling.

(24)	oluguhyo	‘broken pot’
	luukizaano	‘green (Cl. 11)’
	oluguhyó lúúkizaano	‘green broken pot’
	ekikáláángilo	‘frying pan’
	kizito	‘heavy’
	ekikáláángiló kízito	‘heavy frying pan’

As (25) shows, no H is inserted if the following modifier has an H tone.

7 Leftward Spreading creates the only surface counterexamples to this generalization, wherein /i-tí/→[ítí] ‘tree.’

(25) oluguhyo lúno 'this pot' oluguhyo luzímá 'good pot'

The interaction between Lapse Avoidance and MR is seen in (26), where the noun has an underlying penult H. Under a derivational account, these data show either that Lapse Avoidance must be prevented from applying when the target vowel is preceded by a H (i.e. is subject to OCP blockage), or that the H inserted by Lapse Avoidance is subsequently deleted by MR.

(26)	ihéénze	'cockroach'	ihéénzé lyaangu	'quick cockroach'
	ebhalúúwa	'letter'	ebhalúúwáá ndeehi	'long letter'

If H were inserted on the final vowel, it would spread rightward giving incorrect **ihéénzé lyáangu*.

In the OT account, these data are explicable given a specific interpretation of the two-level condition on adjacent Hs. Underlyingly adjacent H tone sequences are banned, and HH sequences which are not underlyingly adjacent are tolerated. These data show that the two-level condition against HH does not care about the underlying status of the second H. The constraint */H/H is thus not a constraint against underlyingly adjacent Hs, which would be too broad a statement, but is, specifically, a constraint against surface H immediately preceded by an underlyingly H toned TBU. MR is sensitive to the underlying status of only the first TBU in an HH sequence, and the constraint considers only the surface status of the second TBU. If the TBUs are underlyingly adjacent (as they are in (26)), and if the first TBU has an underlying H (as it does in (26)), then the constraint prohibits surface HH, whether the second surface H also is an underlying H (as in the examples of (20), (22), (23)) or an inserted H (as in (26)).

So far, the two-level account of MR has fared well enough: problematic data will now be considered. In the data considered so far, an H on a TBU blocks a following H only if the leftmost TBU is underlyingly H. This is not always the case: some cases of rightward tone shift feed into MR. Like all object prefixes in the language, the 1sg object prefix has an underlying H tone: it is underlyingly a moraic nasal, which cannot bear tone on the surface. Consequently, the H from the nasal shifts to the following syllable.[8] The examples below illustrate this shift before underlyingly toneless stems. In the first example, the H shifts to the final syllable, and because that syllable is the utterance final syllable, the

[8] The nasal's mora transfers to the preceding vowel, causing compensatory lengthening of the underlyingly short vowel.

H spreads to the left. In the second example, the H shifts to the penult, but Tone Doubling is prevented from spreading that H to the final syllable. In the third example, H shifts from the nasal to the stem-initial syllable, and then undergoes Tone Doubling.

(27)	'to V'	'to V us'	'to V me'	
	kusya	ku-tú-sya	kúú-n-syá	'grind'
	kubala	ku-tú-bála	kuu-m-bála	'count'
	kutwaangila	ku-tú-twáángila	kuu-n-twáángíla	'pound for'

As the data in (28) show, when the following stem begins with an H tone, the tone pattern is the same. This can be explained either by assuming that MR causes the root H to delete after the object prefix's H, and then H shifts from the prefix, or by assuming that the prefix H shifts to the H-toned root-initial syllable, thus merging with that H.

(28)	'to V'	'to V us'	'to V me'	
	kúlyá	ku-tú-lya	kúú-n-dyá	'eat'
	kubóna	ku-tú-bóna	kuu-m-bóna	'see'
	kutéékéla	ku-tú-téékela	kuu-n-téékéla	'cook for'

Examples such as *kuumbóna, kuuntéékéla* which derive from /ku-ḿ-bóna/, /ku-ń-téékela/ pose a problem for the two-level analysis of MR. The two-level account was crucial in forcing overzealous deletion of H (whereby /abatáá-tú-gí-kálaangiizye/ becomes *abatáá-tú-gi-kalaangiizye* and not **abatáá-tú-gi-káláángiizye* by a more conservative pattern of H deletion). The essential contribution of the two-level constraint is that the tone borne by the first syllable in the sequence should be underlyingly H, both underlyingly present and underlyingly on the first TBU in the HH sequence, without reference to the surface tone. However, this statement of the constraint incorrectly predicts that no H tone at all should surface in /ku-ń-téékela/.

(29)

		ku-ń-téékela	(nasal tonotactics)	*/H/H	Max-IO(H)
☜	a.	kuuntéékéla		*!	*
💣	b.	*kuunteekela			**
	c.	kuńtéékela	*!		*

Compare the analogous choice involving the 1pl object prefix *tú*:

(30)

	ku-tú-téékela	*/H/H	Max-IO(H)
☞	kutútéékela		*
	kututeekela		**!
	kututéékéla	*!	*

Deletion of all Hs is blocked in (30) as being gratuitous: deletion of a single H suffices to avoid violation of MR, and the two-level condition on MR dictates which of the two Hs (the leftmost) will survive. With /ku-ń-téékela/ in (29), H tone cannot be preserved on the leftmost underlying TBU because of inviolable surface tonotactics. Since the actual form (29a) violates MR (*ń* is underlyingly H toned), that form would be wrongly ruled out in favor of the toneless candidate which only violates the relatively low-ranked constraint Max-IO(H).[9] Thus the two-level account which was crucial to explaining the pattern of non-minimal H deletion and the counterfeeding pattern of certain cases of tonal movement is shown to be inconsistent with other facts of the language.

These data pose further problems for an OT analysis. Unlike the situation with H that shifts due to the general Onsetless Tone Shift rule, shift from the tone-bearing nasal of an object prefix feeds into MR. In the phrasal context where H is inserted before a toneless modifier, the H inserted by Lapse Avoidance *is* deleted if the preceding syllable has an H tone that shifts to that position from an H-toned nasal.

(31) ku-bala 'to count'
kuu-m-bála 'to count me; counting me'
kuu-m-bálá kwaako 'your (act of) counting me'

One would expect an H tone to be assigned to the final vowel by Lapse Avoidance—cf. /ku-bala kwaako/ → *kubalá kwáako* 'your counting me'. The reason that no H surfaces on the final vowel is that the preceding vowel *á* has an H tone. However, that vowel does not have an underlying H, but rather has H as a result of rightward tone shift from a tone-bearing nasal. Therefore, it is wrong to say that a TBU must be underlyingly H toned to be visible to MR.

This problem can be resolved in OT if one adopts a partially derivational, multi-modular version of the theory (as has been suggested in various places

[9] One might attempt to avoid this by ranking Max-IO(H) above MR, but such a move can be ruled out by the simple fact that such a ranking wrongly predicts that there is no OCP-driven H deletion at all.

such as McCarthy and Prince 1993; Kenstowicz 1994*a*; Myers 1997; Kiparsky 1999; Itō and Mester 1999), where there are mutiple derivational levels, such as distinct word-level and phrase-level phonologies or even word-internal levels as there are in the theory of Lexical Phonology, and the output of the word-level phonology defines the input to the phrase-level phonology. Under that assumption, the output of the word-level phonology, given underlying /ku-ḿ-bala/, would be *kuumbála*. When this form is resubmitted to the phrasal phonology in the phrase /kuumbála kwaako/, the fact that the H originated on the preceding TBU /ḿ/ in the word-level phonology is inaccessible information. Thus a candidate where H is inserted before the toneless modifier (which would surface as **kuumbálá kwáako* given the constraints that bring about tone doubling) would not be immune to the effects of MR.

There is another line of argument showing that MR is cyclic in the sense of applying distinctly at the word and phrasal levels, with rules interspersed between applications of MR, centering around the fact that MR deletes H on a vowel, and the vowel can then be re-assigned H at the phrasal level. At the word level, MR is responsible for deletion of all but the first in a sequence of underlying Hs (and that leftmost H spreads by Tone Doubling).

(32) kú-há ← ku-há to 'give' (ku-há Búlemo 'to give Bulemo')
ku - gi - tu - ha kugítúha 'to give it to us'
H H → Ø H → Ø

Although /há/ is underlyingly H toned, the H of the preceding object prefix *tú* triggers deletion of the H of /há/. The H of *tú* is itself deleted because it is preceded by *gí* which has H.

Now consider what happens when this word is in a phrase, and Lapse Avoidance becomes relevant. As (33) shows, H tone is assigned to the final vowel.

(33) /ku-gí-tú-há kwaako/ → kugítú!há kwáako 'your giving it to us'

If, in a derivational analysis, MR did not apply first at the word level, and only applied once in a derivation, after Lapse Avoidance at the phrase level, the wrong form would be derived. Lapse Avoidance would have no effect on underlying /ku-gí-tú-há kwaako/, since the final vowel of the stem already has H. MR would then apply (followed by Tone Doubling), deriving **ku-gí-tú-ha kwaako*. The correct result is derived if MR applies first at the word level, deriving *kugítuha*; at the phrasal level, Lapse Avoidance would give *kugítuhá kwaako* (surface *kugítú!há kwáako* via Doubling).

There are also derived environment effects which in a derivational account argue for cyclicity. First, underived tautomorphemic H sequences are not subject to MR, as in *mu-nóó*ꜜ*ló* 'small (cl. 1)', *omuuntu munóóló wáange* 'my small person', where both syllables of the stem *nóóló* have underlying H.[10] Second, MR only applies at the phrasal level to the output of Lapse Avoidance. The fact that MR does not apply to the output of word-internal syllable fusions (/ti-bá-ku-yásama/→*tibákwáásáma* 'they aren't opening the mouth') or tone shifts (/ákuchúmita/→*akúchúmíta* 'he who is stabbing') is explained by ordering MR after these processes at the word level. Ordering cannot explain why these forms do not undergo MR at the phrasal level: the input to the phrasal phonology should be *tibákwáásama*, *akúchúmita*, which satisfies the structural description of MR. However, from the perspective of the phrasal phonology, these forms contain underived HH sequences—HH sequences which are present in the input—and therefore if MR is a cyclic rule and is only applicable to derived forms, phrasal application of MR is automatically blocked.[11]

At the phrasal level, MR only applies if the conditions for the rule derive by application of a rule, and therefore MR only applies in some substring if two conditions are met: the requisite structure (HH) is present, and some rule creates that structure. Insofar as rule application results in increased unfaithfulness to the input, this is equivalent to the condition that the sequence HH is present and there is an IO faithfulness violation with respect to H tone. In OT—following Lubowicz 1998—this translates into conjoining MR with the constraint Dep-H, giving the compound condition "do not both insert H and violate MR." The adjacent Hs of *akúchúmíta* at the phrasal level are tolerated, despite violation of MR, since simple violation of MR is of no consequence, just as the violation of Dep-H which results from the mapping from /oluguhyo

[10] The downstep in the citation form is due to a principle of phonetic interpretation lowering a prepausal H. Since the Hs of the syllables *nóó* and *ló* are distinct Hs, lowering only affects the final syllable. In contrast, in *kúlyá* 'to eat' derived from underlying /ku-lyá/ by Leftward Spread, there is only a single H associated to the last two syllables, hence the two syllables have the same pitch level.

[11] Two empirical questions regarding process interaction cannot be resolved, due to accidental gaps created by the morphology in the language: can onsetless shift feed into MR at the phrasal level, and does shift of H from the object prefix /ń/ feed into MR at the word level? Phrasal H insertion only applies to nouns, but nouns cannot have the H-toned subject prefixes which undergo Onsetless Tone Shift, so the conditions for these two processes never coincide. As for shift of H from /ń/ and word-level MR, the crucial test case would be one where /ń/ immediately precedes a toneless syllable that is itself followed by an H, i.e. something of the form /…ńcvcv́…/. This can only arise in Kikerewe if the second H were the grammatical H tone assigned to the penult or final vowel in certain tenses (see Odden 1998). However, a separate principle deletes all Hs in a word which come before this grammatical H, even those not adjacent to the grammatical H. Accordingly, /ba-laa-ń-balíla/ surfaces as *balaambalíla* 'they will count for me,' just as /ba-laa-ń-hanaantukíla/ becomes *balaampanaantukíla* 'they will descend for me': the potential HH sequence is thus avoided by a separate, even more general mechanism.

lúukizaano/ to *oluguhyó lúúkizaano* is of no importance. What is important in the OT analysis is that these two constraints cannot both be violated in the same substring, as is potentially the case with /kuumbála kwaako/.

It is now time to take stock of the attempt to account for Kikerewe's pattern of OCP-driven H deletion eschewing sequential derivations. A purely non-derivational explanation has failed on two accounts. First, it has proven impossible to come up with any coherent account of the patterns found at the word level, pertaining to non-minimal deletion of H and the interaction between MR and various tone movement processes, even with two-level constraints. Second, it has proven necessary to posit at least two derivational steps in the form of a word-level derivation followed by a phrase-level derivation.

3.3.2 *Constraint conjunction*

The next bit of useful machinery for reconstructing derivations is constraint conjunction. Three such arguments will be given here, one showing how constraint conjunction handles the problematic interaction between two tone sandhi rules in Zinza (another Bantu language of Tanzania), a second showing how constraint conjunction is crucial in handling the morphophonemics of N+C in Kimatuumbi, a Bantu language of Tanzania, and a third involving the interaction between OCP deletion, tone docking and tone throwback in Tachoni, a Bantu language of Kenya.

3.3.2.1 *Zinza Tone* First we consider the problem of Zinza tone sandhi. The essence of the problem is that there are two rules which are in a mutually counterfeeding relation, and no matter how the rules are ordered, the output of the second rule could feed into the other rule, and an OT account is at pains to explain this surface opacity. By the first rule seen in (34), any H tone deletes in a verb if it is followed by an object within the phrase.

(34)	akalima	'he cultivated'	akalima Géeta	'he cultivated in G'
	akatéeka	'he cooked'	akateeka Géeta	'he cooked in G'
	akamúlimila	'he cultivated for him'	akamulimila Géeta	'he cultivated for him in G'

By a second rule, illustrated in (35), any otherwise toneless phrasal head is assigned a final H if it is followed by a toneless complement.

(35)	akalima	'he cultivated'
	akalimá Seengelema	'he cultivated in S'
	$H \rightarrow \varnothing / [_X max[\ldots \_\_ \ldots] Y]$	Insert H on toneless word before toneless word.

The interaction between these processes is seen in (36), where we can see that H deletion deletes any Hs in the verb, which creates a toneless word as an intermediate stage, but because the following word is toneless, an H must then be added at the end of the verb.

(36)	akatéeka	'he cooked'	
	akateeká Seengelema	'he cooked in S'	
	akamúlimila	'he cultivated for him'	
	akamulimilá Seengelema	'he cultivated for him in S'	
	akamúlimila Seengelema		*underlying*
	aka**mu**limila Seengelema		*H deletion*
	akamulimi**lá** Seengelema		*H insertion*

Consider an account of this pattern in OT. We can assume a constraint against H in a word followed by another word in the phrase, "No H-plus," which causes deletion of H in verbs before an object. This constraint dominates faithfulness constraints such as Ident-H which require that underlying tones not be changed. There is also a Lapse Avoidance constraint against two consecutive toneless words in a phrase, which causes insertion of H. Moreover, the H that is inserted under the compulsion of Lapse Avoidance must specifically be assigned to the last vowel of the word, as dictated by a rightward tone alignment constraint.

(37) ***H+:** H tone is disallowed in a phrasal head which is followed by another word.
***L#L:** A toneless word cannot be followed by a toneless word in a phrase.
Align(H,r,ω,r)

Since lexical tones are not shifted or spread to the right in citation forms or non-deleting phrasal contexts, IO faithfulness must dominate the alignment constraint, as seen in (38).

(38)

	akamúlimila	IO-Faith	AR(H)
	akamulimilá	*!*	
☞	akamúlimila		***

Now we come to the tableau in (39), which tries to account for a form that undergoes both deletion of a lexical H and insertion of H at the phrasal level, and the question is, how can we assure selection of the correct form with the lexical H deleted and a final H inserted. The first candidate can be ruled out since it violates the ban against sequences of toneless words. The problem is that the remaining candidates *both* violate the ban on phrase-medial H equally, and while the actually correct candidate better satisfies rightward alignment, we independently know from (38) that rightward alignment is subordinate to IO faithfulness, and therefore it should be better to keep the underlying H tone in position than to delete one H and insert another, or to shift the H to the right, as we have in the actual output.

(39)

	akamúlimila Seengelema	*L#L	*H#+	IO Faith	AR(H)
	*akamulimila Seengelema	*		/ú/	
💣	*akamúlimila Seengelema		*		i i a
☞	akamulimilá Seengelema		*	/ú/ /a/	

(Classical) Sympathy Theory provides no help here. The obvious sympathy candidate would be the intermediate form that you get after H deletion and before phrasal H insertion in a derivational account, with no H tone on the first word. We might identify that sympathy candidate as the best form satisfying No-H-plus, as in (40), although this involves suspending McCarthy's proposal that only faithfulness constraints can be sympathy selectors—see Itō and Mester (1998) and DeLacy (1998), for arguments against limiting the class of selector constraints to faithfulness.

(40)

	akamúlimila Seengelema	*L#L	*H#+	IO Faith	AR(H)
❀*H+	*akamulimila Seengelema	*		/ú/	
	*akamúlimila Seengelema		*		i i a
	akamulimilá Seengelema		*	/ú/ /a/	

But even with this candidate identified, we still have no basis for preferring final H over keeping H in its underlying position. In terms of similarity to the sympathy candidate, the incorrect second candidate and the correct third candidate are *equally* bad, differing only in the *location* of the badness. Given that, it should be left to IO faithfulness to prefer a form with H in the same location as in the input, which leaves us where we started from. This is a kind of 'A to B to A' Duke-of-York derivation which McCarthy points out would pose a problem for Sympathy Theory.[12]

Constraint conjunction (Crowhurst and Hewitt 1998 *inter alii*) can be called on to handle the problem that, of the two candidates which violate the constraint against medial H, we have been unable to discard the more lexically faithful candidate which also violates rightward alignment. But this is just what constraint conjunction is designed to handle, that is, it allows one to pick out from the set of candidates that violate a constraint A, all of those candidates which also violate another constraint B.

(41)

	akamúlimila Seengelema	*L#L	*H#+ ∨ AR(H)	*H#+	IO Faith	AR(H)
	akamulimila Seengelema	*!			/ú/	
	akamúlimila Seengelema		(*) *! (***)	*		i i a
☞	akamulimilá Seengelema		(*)	*	/ú/ /a/	

As the tableau in (41) indicates, ordering the conjunction of No-H-plus and rightward alignment above IO-faithfulness allows us to rule out the candidate which keeps the lexical H in place, but avoids the incorrect implication that there is any *general* inclination for rightward shifting or spreading of tones in the language, and in fact the only directional preference for tone shift, spread, or preservation, happens to be to the left in this language.

It should not be surprising that constraint conjunction could handle some rule ordering. Applying a rule typically results in a pattern of faithfulness violations, but also tends to improve performance with respect to phonotactic constraints. A derivational account involving application of one rule and no application of another rule is thus likely to result in a characteristic pattern of violations and satisfactions of constraints. To the extent that constraint conjunction provides a tool for regulating the acceptance of complex patterns of violations across constraints, it is not hard to see that constraint conjunction is a useful bit of machinery for emulating rule ordering.

[12] See Section 3.4 for discussion of revised Sympathy Theory proposed in McCarthy (1999*d*). Under that theory, and providing that a well-formedness constraint can be the sympathy selector, the Zinza facts succumb to analysis in terms of Sympathy; but then as also shown there, this version of ST does not aparently differ in any significant way from derivational theory in terms of its ability to handle Duke-of-York derivations.

3.3.2.2 *Kimatuumbi NC clusters* A second case where constraint conjunction is crucial in handling ordering comes from Kimatuumbi (a Bantu language of Tanzania (see Odden 1996))—here, neither Sympathy Theory nor two-level constraints will handle the facts. In this language, sequences of nasal plus consonant are subject to different sets of modifications, depending on the derivational source of the sequence. Ultimately, the contrasting effects result from derived differences in the moraicity of the nasal. These effects will be referred to as *mų*-effects and *n*-effects, since the underlying sequences causing these changes are /mų/ for one class of effects, and /nį/ or /ñ/ for the other. One morpheme triggering the *mų*-effect is the class 1 prefix /mų/, whose underlying high back vocoid is seen when attached to a vowel-initial stem or in other contexts where its vowel is lengthened and thus not deletable, as in the first two examples of (42). Otherwise, the vowel /ų/ deletes after /m/. The crucial consonantal change triggered by /mų/ is the nasalization of a following voiced stop, seen in the first example of (b). /mų/ has no effect on a voiceless consonant, and it forms a geminate with a following nasal.

(42)	a. mw-aákį	/mų-áķí/	'hunter'	áka 'to hunt'	
	mųų́-ndų	/mų-n̩dų/	'person'	kaá-ndų	'little person'
	b. m-málaángį	/mų-bálaángį/	'counter'	a-bálaángį pl.	
	m-páandį́	/mų-páandį́/	'planter'	a-páandį́ pl.	
	m-mátį́	/mų-mátį́/	'plasterer'	a-mátį́ pl.	

Another context where the mų-effect can be seen is with the prefix /mų/ marking second plural subjects, whose vowel is optionally deleted, as in (43). Here too we can see nasalization of a voiced stop, in the first example of (b), and no effect on the other consonants.

(43)	a.	mw-aaké		'you (pl.)should hunt'
	b.	m-málaangįte ~	mų-bálaangįte	'you (pl.) counted'
		m-paánde ~	mų-paánde	'you (pl.) should plant'
		n-nóolįte ~	mų-nóolįte	'you (pl.) sharpened'

One morpheme triggering n-effects is the class 9 noun prefix in (44), underlyingly a palatal nasal /ñ/, as seen when the prefix comes before a vowel. When coming before an underlyingly voiceless consonant, this nasal causes voicing, and before another nasal, there is degemination of the nasal. This nasal has no nasalizing effect on a following voiced consonant.

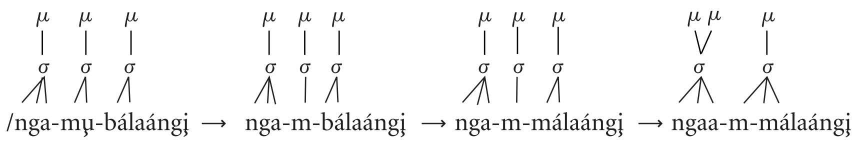
(44)	ñ-epeésį		'light (cl. 9)'
	n-deléká	/ñ-teléká/	'cooked (cl. 9)'
	namátá	/ñ-namátá/	'sticky (cl. 9)'
	m-balaángá	/ñ-balaángá/	'counted (cl. 9)'

A second context illustrating the n-effect involves the 1sg prefix /nį/, where underlying /į/ undergoes optional deletion after /n/; (45) shows that this nasal also triggers voicing and degemination, but causes no nasalization.

(45) n-déljįke ∼ nį-téljįke 'I cooked (recent)'
nóoljte ∼ nį-nóoljte 'I sharpened'
m-bálaangįte ∼ nį-bálaangįte 'I counted'

An important question is what distinguishes the prefixes with respect to changes on a consonant. As argued in Odden (1996), *mų*-prefixes have an intermediate stage where the nasal is moraic, when the nasal effects take place. The mora can be seen in (46) when a vowel precedes the nasal, since the nasal desyllabifies, compensatorily lengthening the preceding vowel.

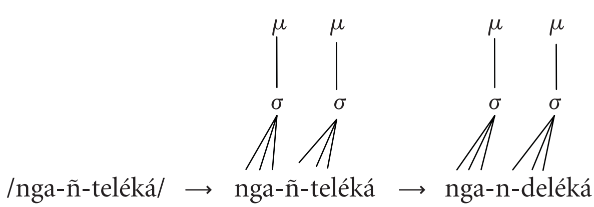
(46)



'it's a counter'
(cf. nga-Lįbųlų́le 'it's Libulule')
kįgómaa m-paánde ←/kįgóma mų-paánde 'you should plant kigoma'
(cf. kįgóma mųpaánde 'you should plant kigoma')

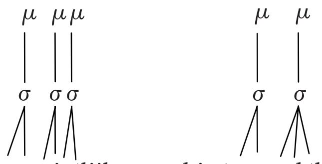
Unlike the *mų*-prefixes, the *n*-prefixes are not moraic at the relevant stage, so the noun prefix /*ñ*/ is simply underlyingly non-moraic. When the class 9 prefix comes before a vowel in (47), there is no lengthening, as there is with other vowel-final prefixes.

(47) ñ-epeésį 'light (cl. 9)'
(cf. mw-eepeésį 'light (cl. 3)')
nga-ndeléká 'it is cooked (cl. 9)'
(cf. ngaa-nteléká 'it is cooked (cl. 3)')



The prefix /nį/ has a vowel, but undergoes vocalic deletion before a consonant without moraic preservation.

(48)



kįgóma nįtélįįke → kįgóma ndélįįke ‘I cooked cassava’

The consequence of mora deletion in the prefix *nį* is that the prefix induces different segmental effects on a following consonant.

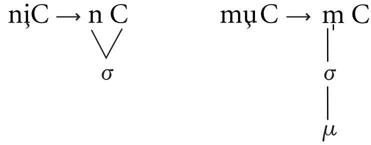
The two classes of nasal plus consonant effects are summarized in (49).

(49)

mųp	→	m̩.p	→	μ + mp		n(į)p→mb
mųb	→	m̩.m	→	μ + mm		n(į)b→mb
mųm	→	m̩.m	→	μ + mm		n(į)m→m

The derivational rules are given in (50).

(50) Vowel deletions: nįC → n C mųC → m̩ C

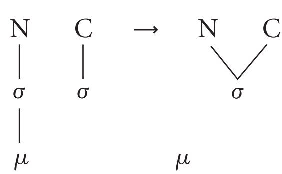


Postnasal voicing: C→ [+voice]/[$_\sigma$[+nas]___

Degemination: C_i → ∅/[$_\sigma$ C_i___

Nasalization: C [+voice]→ [+nasal]/[+nasal] [$_\sigma$___

Nasal readjustment:



The process of vowel deletion affecting /nį/ does not result in a moraic nasal, whereas the deletion affecting /mų/ renders the nasal moraic. Subsequently, various rules are sensitive to these differences in moraicity. Degemination in (51) affects onset geminate nasals, and thus degemination does not affect a moraic nasal plus nasal, since that is not a tautosyllabic cluster.

(51)	*underlying ñ*	*NC derived by deletion of both V and μ*	*NC derived by V deletion with μ preservation*	
	/ñ-nV/	/nį-nV/	/mų-nV /	
	.nnV.	.nį-nV.	.mų.nV.	*Syllabification*
			.m̩.nV.	*V-deletion*
	.nV.	.nV.		*Degemination*
			μ.nnV.	*Nasal resyllabification etc.*

Postnasal onset voicing in (52) only affects onset clusters of nasal plus consonant, and thus does not affect a moraic nasal plus a consonant in a following syllable for the same reason.

(52)	*underlying ñ*	*NC derived by deletion of V and μ*	*NC derived by V deletion with μ preservation*	
	/ñ-tV/	/nį-tV/	/mų-tV/	
	.ntV.	.nį.tV.	.mų.tV.	*Syllabification*
		.ntV.	.m̩.tV.	*V-deletion*
	.ndV.	.ndV.	NA	*Postnasal voicing*
			μ.ntV.	*Nasal resyllabification etc.*

The nasalization process in (53), which affects voiced consonants, affects only a consonant preceded by a moraic nasal, and not an onset nasal plus consonant.

(53)	*underlying ñ*	*NC derived by deletion of V and μ*	*NC derived by V deletion* with μ preservation	
	/ñ-bV/	/nį-bV/	/mų-bV/	
	.mbV.	.nį.bV.	.mų.bV.	*Syllabification*
		.mbV.	.m̩.bV.	*V-deletion*
	NA	NA	.m̩.mV.	*Nasalization*
			μ.mmV.	*Nasal resyllabification etc.*

The opacity of all of these processes is due to the fact that moraic nasals are surface desyllabified, with the mora being transferred to any preceding vowel.

The quandary to be resolved in an OT account is how to distinguish various kinds of nasal plus consonant sequences, depending on their derivational source. We can quickly rule out a two-level account. The basic idea of a two-level account of postnasal voicing would be to penalize sequences of nasal plus consonant, but only if the consonants were underlyingly adjacent. This would correctly allow voicing to be triggered by the noun prefix /ñ/, since it is not followed by a vowel, and would block voicing in the case of underlying /mų/

plus consonant. The problem is that this does not distinguish /mų/, which does not trigger voicing, from /nį/ which does, even though in both cases the consonants are underlyingly nonadjacent.

(54) *N [-voice] (if underlyingly adjacent)

	ñt	*NT	Ident(voi)
	nt	*	
☞	nd		*

	mųt	*NT	Ident(voi)
☞	nt		
	nd		*

	nįt	*NT	Ident(voi)
💣	nt		
	nd		*

The beginnings of a Sympathy-based account are given in (55). In accounting for the opacity of postnasal voicing with respect to deletion of /ų/, the goal would be to identify the intermediate stage, /mųt/, via some failed candidate and require identity to that candidate with respect to voicing. To derive opacity specifically in the case of reduction of /mų/ but not /nį/, we will identify the sympathy-inducing constraint more precisely as Max-round which penalizes deletion of round vowels. Thus deletion of *ų* induces no voicing, despite the phonotactic constraint, because the output is required to look like the intermediate form where no round vowel is deleted.

(55)

	mųt	*mųC	$\text{ID-Voi}_{\text{Max(rd)}}$	*NT	Max(rd)
❀$_{\text{Max(rd)}}$	mųt	*!			
☞	nt			*	*
	nd		*!		*

In the case of underlying /nt/ where no vowel is deleted in (56), the best candidate satisfying Max-round is simply the phonotactically best candidate, which avoids voiceless consonants after nasals.

(56)

	ñt	*nįC	ID-Voi$_{Max(rd)}$	*NT	Max(rd)
	nt			*!	
☞ ❀$_{Max(rd)}$	nd				

And when the deleted vowel is /i/ in (57), the flower candidate is also the phonotactically best form, since although a vowel is deleted in that form, it is not a round vowel.

(57)

	nįt	*nįC	ID-Voi$_{Max(rd)}$	*NT	Max(rd)	ID(voi)
	nįt	*!				
	nt		*!	*		
☞ ❀$_{Max(rd)}$	nd					*

As indicated in the tableaus of (58), the same kind of analysis will account for the preservation of geminate nasals arising from deletion of /ų/, in contrast to degemination as found with underlying nasal plus nasal sequences, or nį plus nasal sequences.

(58)

	mųn	*mųC	Max-C$_{Max(rd)}$	*NN	Max(rd)
❀$_{Max(rd)}$	mųm	*!			
☞	nn			*	*
	n		*!		*

	nn	*nįC	Max-C$_{Max(rd)}$	*NN	Max(rd)	Max(C)
	nn			*!		
☞ ❀$_{Max(rd)}$	n					*

	nįn	*nįC	Max-C$_{Max(rd)}$	*NN	Max(rd)	Max(C)
	nįn	*!				
	nn			*!		
☞ ❀ $_{Max(rd)}$	n					*

We cannot handle the interaction between ų-deletion and nasalization of voiced consonants with Sympathy Theory, in particular, we cannot explain why nasalization affects only the output of ų-deletion. In the derivational account, what explains this pattern is the fact that the trigger is derivedly moraic. Under a Sympathy approach, the sympathy candidate would be one where the onset nasal is moraic, thus triggers nasalization of the following voiced stop. The question is how to identify such a form. We could focus on moraic preservation, and make the sympathy constraint be Max-mora, as in (59), which attempts to derive [aammwéenį] from /a-mų-bwéenį/. But this won't work, since the actually best candidate satisfying Max-mora is the third one in (59) where the nasal itself is *not* moraic, but the mora is nevertheless preserved in the output by being transferred to the previous vowel, so the required sympathy candidate can't be identified: if the sympathetic candidate cannot be identified, there is no basis for allowing nasalization in this case.[13]

(59)

	a-mų-bweenį	Max(μ)	*mųC	*N̩
	amųbweenį μ μ		*!	
(❀)	am̩bweenį μ μ			*!
❀Max(μ)	aambweenį μ μ			
	ambweenį μ	*!		

For Sympathy to work, we have to presume that syllable structure is present underlyingly, and there is a syllable node that underlyingly dominates *mų*. To identify a form with a moraic nasal as the sympathy candidate, the sympathetic constraint will be Max-IO-syllable. The tableau in (60) shows how the flower candidate can be identified. The crucial difference between this approach and the Max-mora approach is that while transfer of the mora to the preceding vowel preserves the mora, it does not preserve the syllable, whereas the candidate where the nasal becomes syllabic also preserves the syllable.

[13] Note incidentally that the candidate also cannot be identified by requiring that the mora be preserved in an IO-faithful manner, since the mora is actually underlyingly on the deleted vowel ų.

(60)

	mų.b σ σ	Max-σ	*mųC	*N̩	*N̩D	Max-V
	m̩.b σ σ			*	*!	*
❀Max-σ	m̩.m σ σ			*		*
	μ mb σ	*!				*
	mų.b σ σ		*!			

Given the right flower candidate so identified, the actual form can be selected because of its similarity to the flower candidate with respect to consonant nasality, as in (61).

(61)

	mų.b σ σ	*mųC	Ident-Nas$_{Max\text{-}\sigma}$	*N̩	*N̩D	Max-V	Ident-Nas
☞	mm σ					*	*
	mb σ		*!		*	*	
❀Max-σ	m̩.m σ σ			*!		*	*
	mų.b σ σ	*!	*				

A theoretical problem with this approach is that it works only if one assumes that inputs in Kimatuumbi are syllabified, at least in any string that can lead to a nasal plus consonant sequence, which contradicts the premise of richness of the base. Moreover, this analysis requires that there actually *be* IO faithfulness constraints for the syllable, which McCarthy (1999*d*) claims is simply not the case. A fatal empirical problem with the sympathy analysis is that it also incorrectly predicts that there should be nasalization when the vowel /į/ deletes, but we know there is no nasalization resulting from reduction of *nį*. Since we are looking for the best form that preserves the underlying syllable of /nį/, we cannot help but find the second candidate in (62a), and prefer it over the first candidate, which then leads us to incorrectly require identity with respect to nasality in deriving the actual form in (62b).

(62)

a.

	nįb	Max-σ	*nįC	*N̩	*N̩D	Max-V
	m̩b			*	*!	*
❀ Max-σ	m̩m			*		*
	µ mb	*!				*
	nįb		*!			

b.

	nįb	*nįC	Ident-Nas$_{Max\text{-}\sigma}$	*N̩	*N̩D	Max-V	Ident-Nas
💣	*mm					*	*
☞	mb		*!		*	*	
❀ Max-σ	m̩m			*!		*	*
	nįb	*!					

There is no way to identify the correct sympathy candidate in this case: somehow, in the case of į-deletion, we have to exclude a sympathy candidate which is, in character, the same as the one that we relied on to derive the correct form in the case of ų-deletion.

There is a way in OT to solve this problem, by appealing to constraint conjunction. The basic idea of the constraint conjunction approach is to say that it is acceptable to delete a round vowel and thus violate Max-rd, or to violate the ban on nasal + voiced sequences, as seen in numerous cases of [nd]. What is not acceptable is to violate both constraints at the same time. By conjoining Max-round with a ban on nasal plus voiced stop, and ordering that conjunction above the relevant faithfulness constraints, we get the desired effect.

(63) (*ND ∨ Max(rd))

The tableaus in (64) show how this works. It is in the case of underlying *mų* in the first tableau that the conjunction is relevant. The serious competition is between the last two candidates, one with nasalization and one without. The candidate without nasalization is out because it has violated both halves of the conjunct. In the second and third tableaus, the conjunction is irrelevant, since no round vowel is being deleted, and thus the conjunction cannot be violated.

(64)

	mųb	(*ND ∨ Max(rd))	*mųC, *nịC	Max(rd)	ID-Nas	*ND
	mųb		*!			
☞	mm	(*)		*	*	
	mb	(*) *! (*)		*		*

	nịb	(*ND ∨ Max(rd))	*mųC, *nịC	Max(rd)	ID-Nas	*ND
	nịb		*!			
	mm				*!	
☞	mb	(*)				*

	nb	(*ND ∨ Max(rd))	*mųC, *nịC	Max(rd)	ID-Nas	*ND
☞	mb					*
	mm				*!	

Thus constraint conjunction is one way in OT to handle data that would otherwise be intractable and would therefore refute OT.

3.3.2.3 *Tachoni Tone* The third case involving crucial use of constraint conjunction in disposing of rule ordering comes from the analysis of tone-mapping principles in Tachoni, a Bantu language spoken in Kenya. The data in (65) illustrate the tone pattern of verbs which are not inflected with the melodic H, a tense-aspect marker of the language. Underlying H-toned vowels are underlined.

(65) *Toneless verbs*

oxu-sy-a	'to grind'	oxu-bal-a	'to count'
oxu-chiing-a	'to carry'	oxu-kaban-a	'to divide'
oxu-karuxasy-a	'to invert'	oxu-chiichakan-a	'to continue'
oxu-yoombool-a	'to spill tr.'	oxu-beechakal-a	'to belch'

H verbs

oxu-bék-a	'to shave'	oxu-téex-a	'to cook'
oxu-búkul-a	'to take'	oxu-fúúndix-a	'to knot'
oxu-bótooxan-a	'to go around'	oxu-ng'ínaang'iny-a	'to shine'
oxu-xámulul-a	'to strain'	oxu-syáánixil-a	'to dry at fire'
oxu-fúkirisany-a	'to agree'	oxu-táángaasy-a	'to announce'

Verbs in this language come in two varieties: H-toned and toneless. If the root is H-toned, the H is realized on the first root syllable, as a level H on a long vowel except in penult position, where it is realized as a falling tone. Somewhat exceptional are monosyllabic H verbs, where the lexical H is realized on the pre-stem syllable.

(66)

oxú-fwa	'to die'	oxú-ha	'to give'
oxú-lya	'to eat'	oxú-nywa	'to drink'
oxú-rya	'to fear'	oxú-ya	'to be ripe'

This leftward shifting of H is due to a principle shifting H off of the final syllable—the number of contexts where H can appear on a final syllable is very small, and can be explicitly enumerated.

Contrasting with this simple pattern is the pattern exhibited when a melodic H tone is added to the verb, as happens in the near future tense. Examples are given in (67). Boxed data indicate surface forms involving complex interaction between phonological principles.

(67) Toneless verbs

ba-li-sya	'they will grind'	ba-li-bála	'they will count'
ba-li-chíínga	'they will carry'	ba-li-kabána	'they will divide'
ba-li-chiingána	'they will carry e.o.'	ba-li-karúxásya	'they will invert'
ba-li-laambáála	'they will lie down'	ba-li-chiichákána	'they will continue'
ba-li-sukúwáníla	'they will scrape for e.o.'		

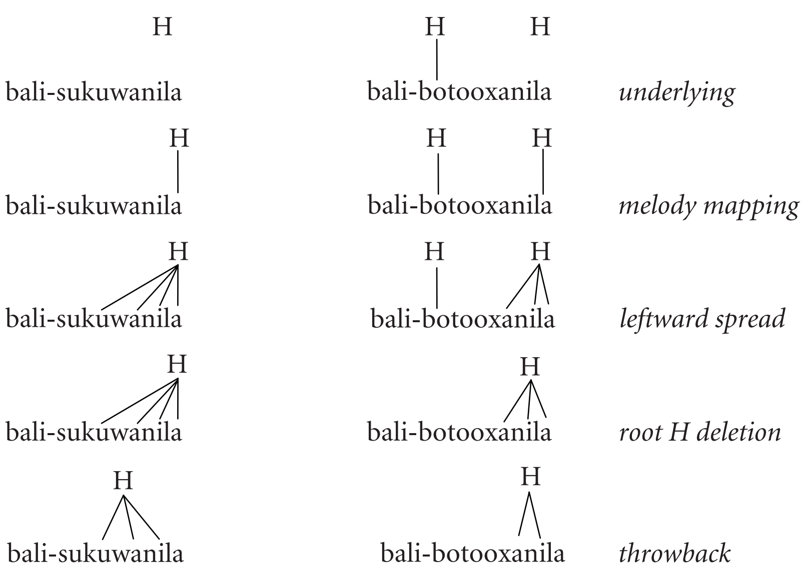
ba-li-lya	'they will eat'	ba-li-beka	'they will shave'
ba-li-teexa	'they will cook'	ba-li-bukúla	'they will take'
ba-li-karáánga	'they will fry'	ba-li-fuundíxa	'they will knot'
ba-li-botooxána	'they will go around'	ba-li-ng'inaang'ínya	'they will knot'
ba-li-xamúlula	'they will strain'	ba-li-syaanixíla	'they will dry at fire'
ba-li-taanagáásya	'they will announce'	ba-li-fukriísánya	'they will agree'
ba-li-botooxáníla	'they will go around for'	ba-li-botooxánáníla	'they will ... for e.o.'

The analysis of this pattern in a derivational account is straightforward. An H tone is assigned to the final vowel as long as that vowel is not also stem-initial, and is later delinked if the preceding syllable has an H tone. Subsequently, the H spreads leftward, stopping at the stem-penitial syllable, or when the preceding syllable has an H tone. Any underlying H in the stem to the left of the melodic H is deleted; finally, H shifts off of the final syllable. The derivation in (68a) illustrates the analysis with a relatively long toneless verb stem, and (68b) does the same for an H stem.

(68)

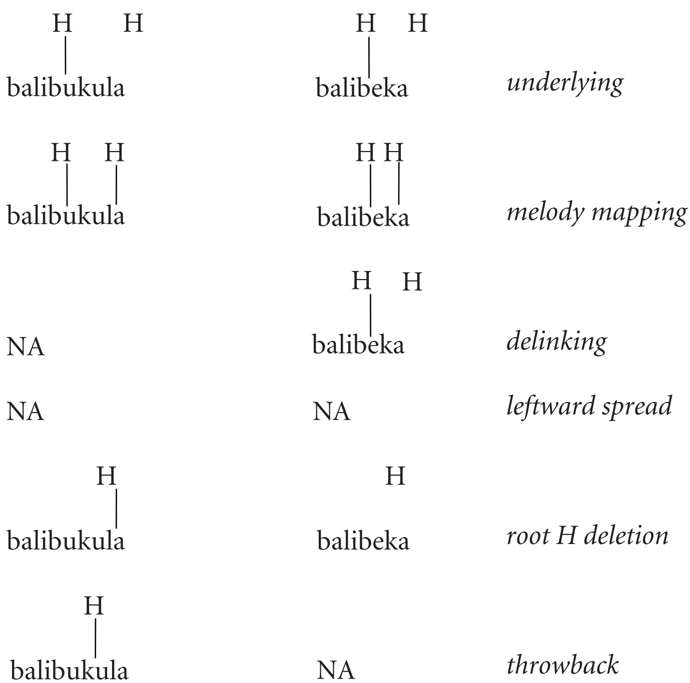
a. ba-li-sukúwáníla
'they will scrape for e.o.'

b. ba-li-botooxáníla
'they will go around for e.o.'



The difference between trisyllabic and disyllabic H roots is made clear in (69), where we can see that while the melodic H is realized on a trisyllabic stem because at the stage in the derivation where the melodic H is first docked to the final vowel, the two Hs in the stem are not on adjacent syllables, whereas in the case of a disyllabic stem, the melodic H is adjacent to the root H and is thus set adrift (or, is prevented from docking in the first place, the choice being empirically insignificant).

(69) a. ba-li-bukúla 'they will take' b. ba-li-beka 'they will shave'



The analysis of these patterns in OT is challenging, and it must answer one question pertaining to the non-surface nature of these generalizations: how does a deleted H block the basic assignment of melodic H, and why doesn't the deleted H also block leftward shifting of final H? The basic analysis of Tachoni, without concern for rule ordering (opacity), can be expressed via five constraints. The melodic H must be mapped to the stem, due to high ranking of Max-H. Insofar as the H is realized on a sequence of vowels, both right and left alignment are active, but since the melodic H is not actually realized on the initial or final syllables themselves, non-initiality and non-finality are also active. Non-finality is only rarely violated, and surface-final Hs can be disregarded (they can be required by specific constraints when they appear). Stem-initial H, on the other hand, is not rare—on the contrary, in lieu of a melodic H, the root H of a verb will appear on the first syllable—and in fact it is only the melodic H which is blocked from initial position in the stem. Therefore, Non-initiality will be restricted to affecting only the melodic H. The tableau in (70) shows how one representative form can be derived.

(70)

	balisukuwanila H	Nonfin	$\text{Noninit}_{\text{Melodic}}$	MaxH	AL	AR	IdentH
a.	balisukuwanila			*!			
b.	balisukuwanilá	*!			uuai		a
c.	balisukuwaníla				uu!a	a	i
d.	balisúkúwáníla		*!			a	uuai
☞ e.	balisukúwáníla				u	a	uai

Turning to an H-toned stem, we also require a constraint prohibiting multiple H tones within the stem, *H . . . H, in order to motivate deletion of the root H.[14] A strictly surface-oriented approach cannot explain why underlyingly H-toned stems block spreading of the melodic H to the second stem syllable in *balibo̲tooxáníla*, not **balibo̲tóóxáníla*; intuitively, this is because there would be an OCP violation, except that the initial syllable is not actually H-toned. As with Kikerewe, we can account for this pattern by positing a two-level version of the OCP, one which prohibits a surface H on a syllable that is after a syllable which is underlyingly H-toned. Armed with such a constraint, the correct form can be derived.

[14] The choice of which tone to delete can be handled by a high-ranking stipulation that melodic tones must be parsed.

(71)

	balibótooxanila H	*/H/H	*H…H	MaxH	AL
a.	balibótooxaníla		*!		oooa
b.	balibotooxaníla			*	ooo!a
☞ c.	balibotooxáníla			*	ooo
d.	balibotóóxáníla	*!		*	o

An alternative is to employ a sympathy constraint, one which preserves the tones of the best candidate not deleting any Hs. Thus the flower candidate *balibótooxáníla*, which is essentially the intermediate form in (68b), has no H on *too*, and therefore the best candidate is one that also has no H on *too*.

(72)

	balibótooxanila H	*H…H	Ident(H)$_{Max(H)}$	*HH	MaxH	AL
a.	balibótooxaníla	*!				oooa
b.	balibótóóxáníla	*!		*		o
❀ c.	balibótooxáníla	*!				ooo
d.	balibotooxaníla		oa!		*	oooa
☞ e.	balibotooxáníla		o		*	ooo
f.	balibotóóxáníla		o oo!	*	*	o

Preservation of the melodic H is, apparently, fairly important in the language, and the relevant constraint must outrank the constraint */H/[H], under the two-level account of the blocking effect of the root-initial H. Notice in (73) that the appearance of H on the penultimate syllable results in violation of */H/[H]; the alternative is to block throwback, or to delete the melodic H. Since the melodic H is retained but thrown back to a position right after the underlying root initial H, Max-H for the melodic H must be rather highly ranked.

(73)

	balibúkula H	Nonfin	MaxH	*/H/[H]	AR	AL
a.	balibukulá	*!	*			aaa
☞ b.	balibukúla		*	*	a	a
c.	balibukula		**!			

Under the Sympathy account, it is crucial that the sympathy candidate allow violation of *HH in order to not violate Non-finality.

(74)

	balibúkula H	*H...H	Nonfin	Symp	*HH	MaxH	AL
	balibúkulá	*!	*				**
❀	balibúkúla	*!			*		*
	balibukula			úú!		**	
☞	balibukúla			ú		*	*

The problem which OT faces in accounting for the facts of Tachoni is that preservation of the melodic H is not completely inviolable: in particular, melodic H is simply deleted in disyllabic H roots (*balibeka* 'they will shave'). Under the two-level account of the opaque effect of the root-initial H, the wrong candidate is chosen—one with a penult H tone.

(75)

	balibéka H	Nonfin	MaxH	Noninit$_{Melodic}$	*/H/[H]	AR	AL
☜ a.	balibeka		*!*				
b.	balibeká	*!	*		*		e
💣 c.	balibéka		*	*		a	
d.	balibéká	*!		*	*		

One might attempt to solve the problem by absolutely preventing the melodic H from ever appearing root-initially, but this will not work, since in fact from a toneless root CVC root, root-initial H is actually possible, viz. *balibála* 'they will count.' Thus non-initiality is a violable constraint.

The problem can be easily identified, and the bad candidate can be ruled out, once one notices that the incorrectly derived form **balibéka* both violates non-initiality and deletes an H tone (the root H), whereas *balibála* only violates non-initiality—that is, one cannot both violate non-initiality and Max-H, a concept expressed in (76) by constraint conjunction.

(76)

	balibéka H	MaxH ∨ $\text{Noninit}_{\text{Melodic}}$	Nonfin	MaxH	$\text{Noninit}_{\text{Melodic}}$	*/H/[H]
☞ a.	balibeka			**		
b.	balibeká		*!	*		*
c.	balibéka	*!		*	*	
d.	balibéká		*!		*	*

Thus, the ability to capture the logical notion "not (A and B)" proves crucial in stating rule-ordering generalizations within OT.

3.3.3 *Abstract operational domains*

Another device to be called on to dispose of derivational concepts is the reified domain, relevant for data from Makonde (a Bantu language of Tanzania and Mozambique). The concept "domain" is a general one applicable to mathematical functions and linguistic operations alike, meaning roughly "the set of things that a rule can apply to." Applied to phonology, "domains" have been construed as abstract constituent structures that are posited to account for restrictions on substrings which do or do not undergo a phonological process. Certain domains have achieved favor (though not universal acceptance), e.g. the syllable or the foot, and are presumably part of a restricted set of universal domains. Here we analyze vowel reduction in Makonde (a Bantu language spoken in Tanzania and Mozambique, discussed in Liphola 1999), where the notion "domain" can resolve problems of rule application in an OT account—in this case the notion of "domain" does not correspond to any motivated phonological constituent, and its sole function is to serve as an instruction to reduce a sequence of mid vowels.

In Makonde, unstressed mid vowels optionally reduce to [a]—stress is regularly on the penultimate syllable. Thus when the vowels of the roots *tot* and *tep* are in the penultimate syllable and are therefore stressed, they cannot be reduced, but when some affix follows the root, the root vowel is unstressed, and reduction of the mid vowel is possible.

(77)	kú-tóót-a	(*kú-táát-a)	'to sew'
	kú-tót-áán-a ~	kú-tát-áán-a	'to sew each other'
	kú-tót-ááng-a ~	kú-tát-ááng-a	'to sew repeatedly'
	kú-téép-a	(*kú-táap-a)	'to bend'
	kú-tép-áán-a ~	kú-táp-áán-a	'to bend from e.o.'
	kú-tép-ááng-a ~	kú-táp-ááng-a	'to bend repeatedly'

The data in (78) further show that the high vowels do not reduce.

(78)

kú-píít-a	'to pass'	kú-púút-a	'to wash'
kú-pít-áán-a	'to pass each other'	kú-pút-áán-a	'to wash each other'
kú-pít-ááng-a	'to pass repeatedly'	kú-pút-ááng-a	'to wash repeatedly'

The motivating force behind this vowel reduction is presumably the markedness constraint against mid vowels. One way to eliminate mid vowels is to raise them to high vowels; this repair strategy can be ruled out by positing that preservation of the specification [-hi] is high-ranked, and therefore the remaining strategy of deleting the vowel's place specification is forced. The fact that only unstressed vowels may reduce is the result of a high-ranking of the constraint requiring faithfulness of stressed syllables to underlying place specifications, and this constraint will not be considered further.

(79) *Mid Max(place) Max(-hi)(mid vowels cannot raise)

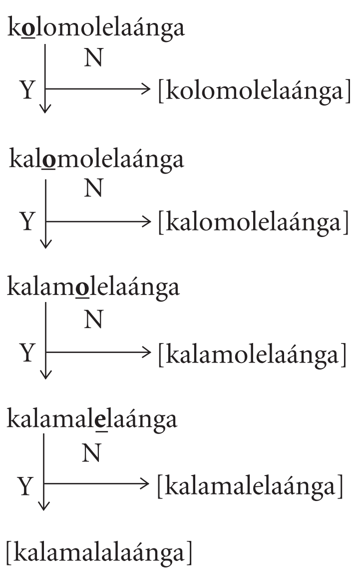
	kútépáána	*Mid	Max(-hi)	Max(place)
	kútépáána	*!		
	kútípáána		*!	
☞	kútápáána			*

While vowel reduction is optional, there is a strict pattern to exercising the option. Reduction begins at the left edge of the stem, and affects any number of vowels, but once reduction has stopped, it is impossible to restart the process. None of the patterns in (b) is possible, since they all involve reducing a vowel after reduction has stopped its left-to-right scan.

(80)	a.	kolomolelaánga	‘cough for (repeated)’
		kalomolelaánga	kalamolelaánga
		kalamalelaánga	kalamalalaánga
	b.	*kolomolalaánga	*kolomalalaánga
		*kolamalalaánga	*kalomalalaánga
		*kalomolalaánga	*kalamolalaánga

This has a simple explanation in derivational theory. Beginning at the leftmost point, one has the option of either applying the rule, or stopping. If at a given stage, the choice is made to stop, this generates a phonetic form where all preceding mid vowels are reduced. The choice to stop can be made at a number of points in the string, thus there are a number of outputs.

(81)



How can this pattern be derived under the assumptions of OT? The first thing to deal with is the fact that for a single input there are many outputs. As (82) makes clear, there is a trading relation between satisfying the ban on mid vowels and satisfying the faithfulness condition on preservation of vowel place. Since any improvement in a form in terms of decreasing the number of mid vowels is paired with loss of place, there is a perfect stalement between these two constraints, as long as they are unranked.

(82)

	kolomolelaánga	*Mid	Max-place
☞	kolomolelaánga	****	
☞	kalomolelaánga	***	*
☞	kalamolelaánga	**	**
☞	kalamalelaánga	*	***
☞	kalamalalaánga		****

However, as (83) also makes clear, there are many other patterns of vowel reduction which result in exactly four stars across the two columns, and not all of these forms are good.

(83)

	kolomolelaánga	*Mid	Max-place
💣	*kolomolalaánga	***	*
💣	*kolomalalaánga	**	**
💣	*kolamalalaánga	*	***
💣	*kalomolalaánga	**	**
💣	*kalomalalaánga	*	***

To resolve this, we can attack the problem structurally by constructing a kind of abstract "domain," especially if we assume a model like Optimal Domains Theory (however, this use of "domain" departs from observed usage in that theory, and should not be taken to imply that this *is* an analysis within ODT). Here, the function of the domain is simply to be a diacritic structure wherein mid vowels are required to reduce, via the constraint $\text{*Mid}_{\text{r-d}}$ which prohibits mid vowels within the R-D constituent. By tying the occurrence of reduction to the structure of this domain, and by judiciously constraining the edges of the domain, we can derive the observed pattern of vowel reduction. As spelled out in (84), the domain is absolutely aligned to the left edge of the stem, and the right end of the domain can be any position after that. The pattern of optionality in reduction then reduces to different sizes of reduction domain, each of which is equally good.

(84)

		kolomolelaánga	A(r-d,l,stem,l)	*Mid$_{r-d}$	A(r-d,r,stem,l)	A(r-d,r,stem,r)
☞	a.	()kolomolelaánga				kolomolelaanga
☞		(ka)lomolelaánga			ka	lomolelaanga
☞		(kala)molelaánga			kala	molelaanga
☞		(kalama)lelaánga			kalama	lelaanga
☞		(kalamala)laánga			kalamala	laanga
	b.	kolomo(la)laánga	ko!lomo			
		kolo(mala)laánga	ko!lo			
		ko(lamala)laánga	ko!			
		(ka)lomo(la)laánga	ka!lomo			
		(ka)lo(mala)laánga	ka!lo			
	c.	*(kalomola)laánga		*!*		
		*(kolomola)laánga		*!**		
		*(kolomala)laánga		*!*		
		*(kolomala)laánga		*!		

The candidates in (84a) are all acceptable, since the left edge of the structure is absolutely at the left edge of the stem, and there are no mid vowels within the structure. In the group of bad candidates in (b), reduction follows non-reduction, and these forms can be ruled out because a domain structure is not left-aligned with the left edge of the stem. In the final group of bad candidates, in (c), the reduction structure is perfectly aligned to the left edge of the stem, but not all mid vowels within the domain are reduced. By judicious use of such a structure, one can handle the problem posed by the pattern of iteration found in Makonde vowel reduction. The question to be asked is whether it is a good thing to add such devices as process-triggering domains to the arsenal, when they have no independent justification or function in the language and do not correspond to phonological entities justified in other languages.

3.4 Sympathy and Duke-of-York Derivations

McCarthy (1997*b*) claims that Sympathy Theory is more restrictive in precluding Duke-of-York derivations of the form A→B→A, where a form is first changed, and later the changed property is restored. In a modification of Classical Sympathy intended to address a counterexample to this claim, McCarthy (1999*d*) sets forth Extended Sympathy Theory, introducing the concept of "Cumulativity," which is the requirement that the sympathetically successful output must accumulate all of the IO faithfulness violations of the sympathy candidate. This section revisits the data from Zinza discussed in Section 3.3.2.1 and explores the issue of inheritance of faithfulness violations, with respect to the derivation of the form /akanywá Seengelema/ → *akanywa Seengelema* → [akanywá Seengelema]. We will see that EST is identical to derivational theory in important ways in its ability to reconstruct such derivations.

McCarthy (1997*b*, 1999*d*) emphasizes the supposed impossibility of so-called Duke-of-York derivations in OT. Suppose that we have a derivation such as (85).

(85) /ABC/
ADC (B → D / __ C: Underlying /B/ can condition the next rule)
EDC (A → E / __ D: The rule applies)
EBC (D → B / E__: Then the segment is turned back into a B)

The problem is explaining why /A/ changes to [E] when the motivating segment, [D], is not actually found on the surface. With Sympathy Theory, a solution might be essayed by making reference to a sympathetic candidate *EDC*, where the trigger segment [D] is actually found. But to be identified as the sympathy candidate, *EDC* must be the best candidate satisfying the selector constraint, which McCarthy stipulates must be an IO Faithfulness constraint; therefore this supposed sympathy candidate *EDC* must be more like *ABC* than another imaginable sympathy candidate *EBC*. But obviously *EBC* would be more faithful to the input than *EDC*, and therefore *EDC* could *not* be identified as the sympathy candidate. To the extent that the sympathy candidate is a reconstruction of an intermediate derivational stage, the sympathy candidate and therefore the intermediate form *EDC* cannot exist, and DY derivations are impossible in OT.

Kiparsky (1999) shows that some DY derivations *are* theoretically possible in Sympathy Theory, and reconstructs a hypothetical Duke-of-York derivation in OT, using Sympathy Theory. The derivation is given in (86). The crux of this example is that the winning candidate is fairly similar to the sympathy candidate, and is actually closer to the input, in terms of IO faithfulness, than the sympathy candidate is.

(86) /maat/

maati	Epenthesis (repair trimoraic syllable)
maači	Palatalization
maač	Final apocope
mač	Shortening (another trimoraic syllable repair)

As seen in (87), from a sympathy candidate *maači*, the surface form [mač] can be selected over the competitor **mat* by requiring consonantal identity with the sympathy form. It is a simple matter to identify *maači*, since that is the best candidate which preserves all input moras.

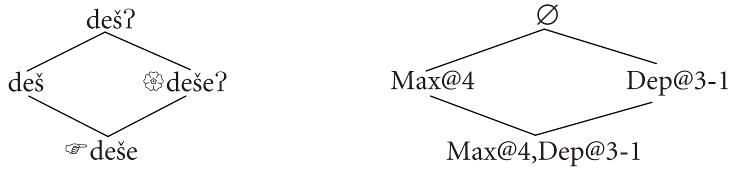
(87)

maat	$^*\mu\mu\mu]_\sigma$	*ti	❀ Ident(hi)	Ident(hi)	Dep-V	☆Max-μ
mat			*!			*
❀ maači				*	*!	✓
☞ mač				*		*
maat	*!		*!			✓
maati		*!	*!		*	✓
maač	*!			*		✓

While no language is given with this property, this demonstrates that DY derivations are not entirely out of the reach of OT.

In response to this challenge to the claim of restrictiveness, McCarthy (1999*d*) modifies ST based on the concept of "cumulativity of unfaithfulness." In this revision, a critical prerequisite for being judged successful in terms of similarity to a sympathy candidate is that every faithfulness violation of the sympathy candidate must also be found in the winning candidate. This requires not just keeping track of, for example, how many Ident-IO violations there are in a candidate, but also the exact location of the violations. The classical case of Tiberian Hebrew /dešʔ/→ [deše] is graphically represented as in (88).

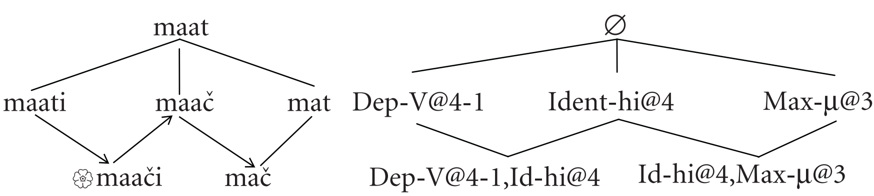
(88)



Here, *deš* has violated Max at input segment 4 (by deleting the glottal stop), and the flower candidate *deše?* has violated Dep by inserting *e* to the right of segment number 3 (notated as Dep @3-1). The actual output form *deše* inherits the Dep @3-1 violation of *deše?* plus adds the violation Max @4: in contrast, the candidate *deš* fails to inherit the violation Dep @3-1 from the sympathy candidate, and thus is excluded. That is, *deš* does not accumulate the unfaithfulness of the flower candidate *deše?*, because it does not have the unfaithful epenthetic vowel.

Turning to the attempted DY derivation of *mač* from /maat/, (89) shows that the desired output candidate *mač* has the faithfulness violation Ident-hi @4 in common with the sympathetic candidate, plus the violation Max-μ @3. However, neither the desired output nor the near competitor *mat* inherit from *maači* the violation Dep-V @4-1, that is, it does not retain the epenthetic vowel, and thus in McCarthy's terms these candidates are *non-comparable* to the flower candidate.

(89)



Since neither *mač* nor *mat* have inherited all of the faithfulness violations of *maači* (lacking the epenthetic vowel), *mač* and *mat* cannot be distinguished in terms of faithfulness to the sympathy candidate, and indeed all candidates other than the flower candidate itself fails sympathy. Thus sympathy cannot be called upon to emulate this DY derivation, it is claimed.

3.4.1 *The DY derivation of Zinza*

We now return to Zinza tone sandhi, which involves the interaction of a rule deleting phrase-medial H in (90a), and a rule adding H to the end of a toneless word before a toneless word in (90b), with the interaction of these processes seen in (90c).

(90) a. akatéeka Géeta → akateeka Géeta 'he cooked in G'
b. akalima Seengelema → akalimá Seengelema 'he cultivated in S'
c. akatéeka Seengelema → akateeká Seengelema 'he cooked in S'
akamúlimila Seengelema → akamulimilá Seengelema
'he cooked for him in Sengerema'

We have seen in 3.3.2.1 that these data cannot be accounted for with Classical Sympathy Theory. Extended Sympathy Theory, on the other hand, can capture the relevant distinction, since the actual output has a property in common with the flower candidate, namely the specific loss of input H on the prefix *mu*.

(91) /$a_1k_2a_3m_4ú_5l_6i_7m_8i_9l_{10}a_{11}$ Seengelema/

	akamúlimila Seengelema	*L#L	❋*H#+	❀ Sym	IO Faith	AR(H)
❀*H+	*akamulimila Seengelema	*!			/ú-/	
	*akamúlimila Seengelema		*	$*\aleph_0!$[15]		i i a
☞	akamulimilá Seengelema		*	{Dep_{11}}	/ú/ /a/	

IO-faithful **akamúlimila Seengelema* is rejected by the sympathy constraint, precisely because it *is* IO-faithful, in not sharing the loss of the lexical H tone. Note, incidentally, that the sympathy candidate is the best (only) candidate satisfying the constraint against phrase medial H tone, which indicates, following Itō and Mester (1999) and De Lacy (1998), that not all sympathy selectors are faithfulness constraints.

Now consider the examples in (92), with underlying Hs on the last two syllables. This form would seem to be a problem for a Sympathy account, since the surface H is realized on a syllable with an underlying H, so this would seem to be a retreat in unfaithfulness that is not supposed to be allowed.

[15] The number of stars assigned to a candidate violating the sympathy constraint is, according to McCarthy, greater than the number assigned to the worst "conforming" candidate, i.e. candidate inheriting all of the faithlessness of the flower candidate. Intuitively, that means that an infinite number of stars is assigned, since there is no upper limit on the length of the worst conforming candidate (and each added segment increases the number of stars). It is generously assumed that the length of the worst candidate is countable, in the technical sense, and thus has size $\aleph_0$, but verification of this assumption requires mathematical proof.

(92)	/aka-mú-pá/	→	aka-mú-pa	'he gave him'
	aka-mú-pá bukoko	→	aka-mu-pá bukoko	'he gave him gray spotted bananas'

The problem is clear in tableau (93).

(93)

	akamúpá bukoko	*L#L	❋*H#+	❀Sym	IO Faith	AR(H)
❀ *H+	akamupa bukoko	*!			/ú-/ /á/	
💣	*akámupa bukoko		*	{Dep3}	/ú/ /á/ /á/	u a
	akamúpa bukoko		*	*$\aleph_0$!	/á/	a
☞	akamupá bukoko			*$\aleph_0$!	/ú/	

The phonetically occurring form *akamupá bukoko* should be rejected by the sympathy constraint because it apparently does not share with the sympathy candidate the loss of input H on the final vowel, and this would lead to incorrect selection of **akámupa bukoko* where the inserted H is only slightly misaligned by being inserted on the rightmost vowel which doesn't have an underlying H tone. A bit more analysis shows that *akamupá bukoko* need not be rejected as constituting a retreat in faithlessness.

Although there is a H on the last vowel of the verb in the output, just as there is in the input, that doesn't mean that the two Hs are the same Hs. If we analyze the final H as being not the retention of the underlying H, but rather the replacement of one underlying H with a different H, then we escape the disastrous sympathy consequences noted in (93). The successful derivation of this form requires distinguishing two phonetically identical candidates, one where the underlying token of H is directly preserved, and another where the underlying H is missing and a different H is inserted.[16] The former candidate incurs the full wrath of violating the sympathy constraint, whereas the latter candidate only suffers the rather mild consequence of adding an H not found in the sympathy candidate.

[16] Specifically, the inserted H has no input correspondent whereas the retained H corresponds to the input H.

(94)

	akamúpá bukoko	*L#L	❃*H#+	❀ Sym	IO Faith	AR(H)
❀*H+	akamupa bukoko	*!			/ú/ /á/	
	akámupa bukoko		*	{Dep$_3$}	/ú/ /á/ /a/!	u a
	akamupa bukoko H_{input}		*	$*\aleph_0$!	/á/	a
☞	akamupa bukoko $H_{inserted}$		*	{Dep$_7$}	/ú/ /á/	

The finger candidate wins in the competition against the one with an antepenultimate H because of differences in IO Faithfulness as well as rightward alignment. This provides a case where a DY derivation is not beyond the reach of OT, any more than it is in derivational theory, and, since such a relation actually exists in Zinza, this would be a desirable result for OT.

Zinza provides independent phonological evidence for distinguishing H tones which are present in underlying representations from Hs which are inserted in response to some constraint, and those phonological tests show that the correct analysis must indeed be one where an underlying H is replaced with an epenthetic H. To see this evidence, we will look at some details of the tone system of Zinza. The essence of the argument is as follows. An H on either of the last two moras of the word spreads bidirectionally to the penult and final moras, as long as the final syllable is not prepausal. However, an inserted H does not spread: then, using the test of tone spreading, the surface H in *akamupá bukoko*, from /akamúpá bukoko/, must be an inserted H tone, not a preserved underlying H tone, since it does not spread to the penult.

3.4.2 *Tone doubling and penult H tones*

As in many Bantu languages, verbs in Zinza may be inflected with a floating melodic tone in certain tenses—see Crowhurst and Hewitt (1998) for an OT analysis of the cognate process in Zezuru Shona, and Poletto (1998) for closely related Runyankore. In Zinza (as in Runyankore and Shona), the position where this H is realized depends on whether the verb stem has an underlying H tone, or is underlyingly toneless. If the verb is toneless, the melodic H is realized on the second mora of the stem (as long as that is not the word-final syllable, in which case the H is realized on the initial mora). The underlying tone of the stem is revealed in the infinitive, and the habitual is one tense exhibiting this melodic H tone.

(95)

ku-lima	'to cultivate'	ba-líma	'they cultivate'
ku-limila	'they cultivate for'	ba-limíla	'they cultivate for'
ku-limiana	'to cultivate for e.o.'	ba-limílana	'they cultivate for e.o.'
ku-libatilana	'they tread for e.o.'	ba-libátilana	'they tread for e.o.'

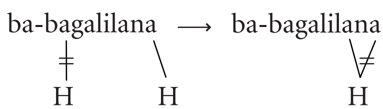
If the stem is underlyingly H-toned, the melodic H appears on the surface penultimate syllable—the underlying H is deleted, since only a single H may appear within the stem.

(96)

ku-bóna	'to see'	ba-bóna	'they see'
ku-bónana	'to see each other'	ba-bonána	'they see e.o.'
ku-témelana	'to chop for e.o.'	ba-temelána	'they see for e.o.'
ku-bágalila	'to weed for'	ba-bagalíla	'they weed for'
ku-bágalilana	'to weed for e.o.'	ba-bagalilána	'they weed for e.o.'

Despite the fact that the H surfaces on the penultimate syllable, it can be argued that the H is assigned to the final syllable and is shifted to the left, by a process shifting any word-final H to the left. Thus the derivation would be as in (97).

(97)



The evidence for assigning the melodic H to the final syllable and shifting it to the left, rather than directly assigning it to the penult, is the fact that H has a different phonetic realization when it is directly assigned to a long penult than it has when a long penult receives H by shifting. An H which is underlyingly on a long penult is realized as a fall. Thus consider the following lexically H-toned roots, where the H is underlyingly on the first syllable of the stem.

(98)

ku-téeka	'cook'
ku-yéela	'go strolling'
ku-kwáata	'touch'

Similarly, if a melodic H tone is assigned to a long syllable, it is realized as a falling tone, as in the following example of toneless CVCVVCV stems in the habitual.

(99)

Infinitive	*3pl. habitual*	
ku-holoota	ba-holóota	'snore'
ku-baziila	ba-bazíila	'sew'
ku-fukaana	ba-fukáana	'wrestle'

In a derivational account, one would assume a rule which shifts an H tone exclusively to the first mora of a long, H-toned penult.

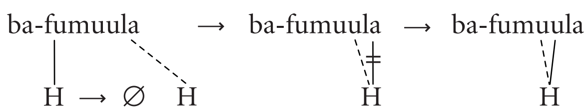
In contrast, an H assigned to the final syllable but retracted to the penult is realized as a level H.

(100)

Infinitive	*3pl. habitual*		
ku-téeka	ba-tééka	←ba-teeká	'cook'
ku-yéela	ba-yééla	←ba-yeelá	'go strolling'
ku-kwáata	ba-kwááta	←ba-kwaatá	'touch'
ku-fúmuula	ba-fumúúla	←ba-fumuulá	'argue'
ku-búúbuuta	ba-buubúúta	←ba-buubuutá	'blow on a fire to ripen bananas'

The melodic H would be assigned to the final syllable, and causes deletion of the preceding lexical H. After application of the rule creating falling tones on H-toned long penults, the final H is shifted to the penultimate syllable, where it is realized as a level H tone.

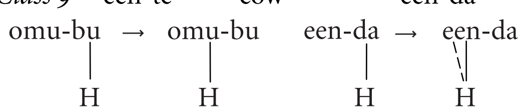
(101)



There is further evidence for a process that shifts final H to the penult. Noun class prefixes are underlyingly toneless, as can be seen in the examples on the left in (102). Just in case the following CV stem is lexically H-toned, the H shifts from the final syllable to the penult: if that syllable is long, the H surfaces as level H.

(102)

Class 3	omu-lilo	'fire'	omú-bu	'mosquito'
Class 5	ii-po	'maize cob'	íí-hwa	'thorn'
Class 6	ama-po	'maize cobs'	amá-hwa	'thorns'
Class 9	een-te	'cow'	één-da	'louse'



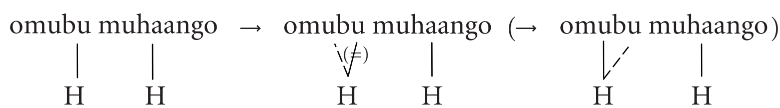
We now turn to the phrase-medial realization of an underlying H tone, either an H lexically linked to the final syllable or a melodic H assigned to the final syllable. The examples in (103) show that the underlyingly final H is realized on the penult and final syllables in utterance-medial position.

(103)

omuuntu	'person'	omúbu	'mosquito'
omuuntu muháango	'large person'	omúbú muháango	'large mosquito'
enyémela	'antelope'	éémbwa	'dog'
enyémelaa mpáango	'large antelope'	éémbwáá mpáango	'large dog'

The analysis of these data is somewhat ambiguous, but the central point is clear, namely that when a word with an underlyingly final H tone is utterance medial, the final H appears on both the final and penult syllables. As we have already seen, word-final H tone shifts to the penult; these data indicate either that shifting should be decomposed into leftward spreading from word-final syllables plus utterance final delinking, or shifting from word-final syllables plus rightward spread from penult to final in utterance medial position.

(104)



A melodic H assigned to the final syllable in an H-toned verb also undergoes this same process. It is necessary to select a verb tense where the phrasal process of H deletion does not apply—the negative habitual is such a tense. The data in (105) are examples of the negative habitual of H-toned verbs.

(105)

tibatééka	'they don't cook'
tibatéékà maláaya	'they don't cook malaaya'
tibabóna	'they don't see'
tibabóná Seengelema	'they don't see Sengerema'
tibasigála	'they don't remain'
tibasigálá Seengelema	'they don't remain in Sengerema'

The data in (106) are analogous in that for reasons of the syntactic context for deleting H in verbs, a verb followed by a postposed subject does not undergo H deletion. These examples that show that phrase final, utterance-medial affirmative verbs which do not lose their H tone similarly spread the final H to the penult.

(106)

atééka	'he cooks'
atééká Bulemo	'Bulemo cooks'
asigála	'he remains'
asigálá Bulemo	'Bulemo remains'

There is further evidence that if an H tone ends up on the penultimate mora, it will spread to the right in utterance-medial context. The perstitive tense is one of those tenses which are not subject to deletion of H tone phrase-medially. As can be seen in the following data, if the verb stem is lexically H-toned, that H appears exclusively on the stem-initial mora, except that if that mora is also the penultimate mora, the H spreads to the final syllable as well.

(107)	tucháá-kúlatila	tucháá-kúlatila chaasa	'we are still following (a chaasa)'
	tucháá-fúlula	tucháá-fúlula maláaya	'we are still transplanting (malaya)'
	tucháá-téeka	tucháá-téeka bukoko	'we are still cooking (bukoko)'
	tucháá-kóma	tucháá-kómá bihógo	'we are still tying (a red cow)'

This provides independent evidence for the process spreading H from penult to final, showing that such alternations are not found exclusively in the context where final H shifts to the penult.

To summarize the details of final and penult H tone in phrase-medial position, we have seen that if an H tone is on an utterance-medial word-final vowel, either because the vowel has an underlying final H, or because the melodic H suffix is assigned to the final vowel, then that H will be realized on the surface on both the final and penultimate syllables. However, these processes do not affect the H tone which is assigned at the phrasal level before a toneless modifier, cf. *kulimá bukoko* 'to cultivate bukoko' (**kulímá bukoko*).

The relevant distinction is not hard to make in OT. In the two cases where word-final H spreads to the penult, the H tone is underlyingly present, either associated to the final vowel (in the case of /omu-bú/ → [omúbu] ∼ [omúbú . . .]), or not associated and just present as a floating H tone suffix, to be mapped to a specfic vowel of the stem by appropriate constraints. In the one case where final H does not spread to the penult, that H tone is not underlyingly present, and is inserted only in response to constraints. In other words, singly linked input-present H tones are disallowed in both the final and penult moras.[17]

[17] It is immaterial whether these two constraints are collapsed into one, or whether abstract foot structure is invoked; it also does not matter whether there is a single phonological principle at work, or two accidentally similar ones.

(108) Non-finality $\begin{array}{c} {}^{*}\mathrm{H_{in}} \\ | \\ \mathrm{V\,V_{\omega}]} \end{array}$ Non-penultimacy $\begin{array}{c} {}^{*}\mathrm{H_{in}} \\ | \\ \mathrm{V\,V_{\omega}]} \end{array}$

These data remind us that a distinction which can be made in OT is between properties found in the input versus ones not found in the input, as encoded in two-level constraints.

Since we now have a diagnostic for distinguishing H tones which are present in the input versus ones inserted in order to satisfy a constraint, we can return to the central question of the DY derivation /aka-lyá bukoko/→ *akalya bukoko*→ [akalyá bukoko] 'he ate bukoko'. This derivation seemed to be problematic, given that there is a word-final H tone in the input, no final H in the intermediate stage (sympathetic candidate), but there is a final H on the final vowel in the output. With our diagnostic for distinguishing input Hs from inserted Hs (i.e. whether the final H spreads to the penult), we can now see that the surface H is not a "restoring" of the input H, but, just like in the derivational account, is a totally separate H. As such, the output form shares with the sympathetic candidate the loss of the input H, and thus does not incur a fatal violation of the sympathy constraint due to non-cumulativity.

(109)

	akalyá bukoko	*L#L	❈*H#+	❁ Sym	IO Faith	AR(H)
	akálya bukoko		*	{Dep_3}	Max_6,Dep_3	a!
☞	akalyá bukoko		*	{Dep_6}	Max_6,Dep_6	
❁	akalya bukoko	*!			Max_6	

Thus at least this kind of DY derivation is not beyond the reach of OT, any more than it is in derivational theory.

3.4.3 *A DY derivation in Kimatuumbi*

Kimatuumbi provides another DY derivation, the crux of which centers around the interaction of three processes, one being a rule that shifts final H tone to a preceding long vowel, one shortening long vowels in a word which is followed by a modifier, and one being Glide Formation, which compensatorily lengthens the following vowel. The interaction between Glide Formation and phrasal shortening is surface-opaque, since long vowels derived by applying GF do not undergo shortening. This might suggest that somehow vowel shortening is blocked from applying just in case Glide Formation has applied.[18] However, tone retraction provides independent evidence that what really happens is that the long vowel is shortened, and is then re-lengthened as a side effect of Glide Formation. The evidence for this is the fact that tone retraction is sensitive to vowel length: final H is retracted only to a long vowel, and when an underlyingly long vowel is shortened, tone retraction no longer takes place. The crucial DY derivation is found in forms such as /mụ-eembé waángu/ which surfaces as [mweembé waángu] 'my mango', in contast to the citation form [mweémbe] 'mango' where shortening does not take place, and final H is retracted. Failure of retraction in [mweembé waángu] can only be explained via the intermediate form [mụ-embé waángu], where there is no long vowel. Thus, /mụ-eembé waángu/→ *mụ-embé waángu* → [mw-eembé waángu].

The first process to be motivated is the phrasal Shortening process (Odden 1987, 1990, 1996) which shortens long vowels in words followed by modifiers illustrated with nouns (110a), verbs (110b), and adjectives (110c).

(110)	a.	kịkól[oo]mbe	'cleaning shell'
		kịkól[o]mbe chaángu	'my cleaning shell'
		mịk[aá]te	'loaves'
		mịk[a]té mịkúlu mịkúlú	'large loaves'
	b.	naan-kál[aa]ng[ịị]le	'I fried for him'
		naan-kál[a]ng[ị]le Mambóondo	'I fried for Mamboondo'
	c.	nn[aá]so	'long (sg.)'
		mịl[a]só mịlaáso	'long (pl.)'

This alternation can be derived by a rule shortening vowels in the head of XP, or via a constraint prohibiting long vowels in the head of XP.

(111) $* \sigma$]X [...]X]^{MAX}
$\mu\ \mu$ (both linked to σ)

The second process is Glide Formation, which desyllabifies a high vowel before a vowel. The data of (112) show this process applying to the combination of a noun class prefix plus a vowel-initial prefix, with the examples on the left showing the underlying vowel before a consonant-initial stem.

[18] Thus constraint conjunction might be invoked, preventing simultaneous violation of those constraints which are characteristic of the application of GF and shortening.

(112)	l[į]-kųn'ųų́nda	'filtered beer'	l[y-oo]wá	'beehive'
	k[į]-kálaango	'frying pan'	k[y-ųų́]lá	'frog'
	[į]-kálaango	'frying pans'	[y-ųų́]lá	'frogs'
	l[ų]-toóndwa	'star'	l[w-aa]té	'banana hand'

Evidence for compensatory lengthening due to Glide Formation is seen in (113), where the form on the left shows the underlying short vowel either word-initial or after the vowel *a*, and the forms on the right showing a long vowel just in case Glide Formation applies.

(113)	[a]té	'banana hands'	lw-[aa]té	'banana hand'
	ka-[ų́]lá	'small frog'	ky-[ųų́]lá	'frog'
	[į]pukú	'rats'	tw-[įį́]pukú	'little rats'

The data in (114) further show that stems may have underlying initial long vowels, and that the underlying long/short distinction is neutralized when Glide Formation applies.

(114)	[eé]mbe	'mango fruit'	mw-[eé]mbe	'mango tree'
	ma-[ée]ke	'storage structures'	ly-[ée]ke	'storage structure'
	[eé]la	'money'	mw-[eé]la	'in money'

Since Glide Formation creates long vowels, and Shortening shortens vowels, we want to know how these processes interact. (115) shows that Shortening does not apply to the output of Glide Formation.

(115)	ly-[oo]wá lįnaántopá	'heavy beehive'	← /lį-owá lįnaántopá/
	ky-[ųų́]la chaángu	'my frog'	←/kį-ų́la chaángu/

This is explained derivationally by ordering Shortening before Glide Formation. In OT, this can be handled by conjoining Max-μ and Ident-μ, the idea being that Glide Formation changes the moraic identity of the prevocalic high vowel (thus signals application of GF), and Shortening results in violation of Max-μ, so that the statement "do not apply GF and then Shortening" translates into the conjunction "do not violate Max-μ and Ident-μ."[19]

The process of Heavy Retraction can be seen at work in the data of (116a–b). These examples are verbs in the subjunctive, where an H tone is assigned to the third stem mora, as seen in (a). The examples in (b) illustrate the case where the third mora is word-final and is also preceded by a long vowel.

[19] This approach actually does not work so simply, since Shortening can apply to the output of Glide Formation, as in the case of /ák-į-an-a įtúumbili/ → *ák-y-aan-a įtúumbili* →*ák-y-an-a įtúumbili* 'to net-hunt monkeys for each other'. The problem can be resolved by applying Glide Formation cyclically, in which case the derived vowel length in the case of *ákyaana* would already be in place when phrasal shortening is encountered. We will disregard this problem here.

(116)

a.	ų-lyé	'you should eat'	ba-temé	'they should chop'
	n-teleké	'you (pl.) should cook'	ų-lindiíle	'you should guard'
	į-n'alan'áate	'it should shine'	ų-bųųndáye	'you should blunt'
b.	ų-kaáte	'you should cut'	ų-toóle	'you should take'

The appearance of H on the second mora, rather than the expected third, can be explained by assigning the H to the third mora, as expected, resulting in a final H (viz. intermediate *ųkaaté*), and then retracting that H to the preceding long vowel, via a rule of Heavy Retraction. In addition to explaining alternations such as those in (116), Heavy Retraction explains why a final H tone cannot (generally) be preceded by a long vowel.

Further evidence for Heavy Retraction, and data demonstrating the interaction between Shortening and Heavy Retraction, can be found in (117). These nouns have a penultimate rising tone—a surface anomaly since generally long vowels with H tone in nouns have falling tone and not rising tone. We can see from the forms on the right that the H tone is underlyingly on the final syllable, and that it shifts to the left just in case it is preceded by a long vowel. If, however, the vowel is shortened because of phrasal Shortening, then the final H remains in its orginal position.

(117)	mboópo	'machete'	mbopó yaángu	'my machete'
	makoóndį	'fists'	makondį́ átatų	'three fists'
	eémbe	'mango fruit'	embé yaangu	'my mango fruit'

Thus, [mboópo] derives from /mboopó/ via Heavy Retraction; phrase-medially, /mboopó yaangu/ undergoes shortening, which bleeds Heavy Retraction.

A long vowel which is created by Glide Formation does not trigger application of Heavy Retraction. This is shown by the examples of (118), where the form on the left presents surface failure of Retraction, and the form on the right motivates the underlying short vowel.

(118)	ly-oowá ← /lį-owá/	'beehive'	ma-owá	'beehives'
	ky-iikį́ ← /kį-ikí/	'stump'	ka-ikí	'little stump'

Now we come to the three-way interaction between Heavy Retraction, Glide Formation, and Shortening. We know that Shortening precedes Glide Formation from the derivation /kį-ų́la chaángu/ → [kyųų́la chaángu], where the long vowel derived by GF does not get shortened. Shortening must precedes Heavy Retraction, because retraction does not take place in /eembé yaángu/→ [embé yaángu] 'my mango fruit' where Shortening has applied. Heavy Retraction must precede Glide Formation because a long vowel created by the latter rule does not trigger retraction, as shown by /lįowá/→ [lyoowa] 'beehive'. From this we derive the strict ordering Shortening > > Heavy Retraction > > Glide Formation. This ordering is directly justified by the derivation of the data in (119).

(119) mweémbe ← /mų-eembé/ 'mango tree'
mw-eembé waángu 'my mango tree'
kyaáme ← /kį-aamé/ 'deserted place'
kyaamé chaángu 'my deserted place'

Beginning with underlying /mų-eembé waángu/, the first rule to apply is Shortening which derives the intermediate form *mų-embé waángu*. At this point, Heavy Retraction cannot apply; then the last rule to apply is Glide Formation, which has the consequence of re-lengthening the vowel, giving the surface form *mweembé waángu*. These data show that a Duke-of-York derivation is crucial to explaining the interaction of processes in Kimatuumbi. It would be insufficient to simply block Shortening from applying in a context where Glide Formation would also apply (the tack taken by McCarthy handling the interaction between dorsal-rounding and final unrounding in Makah). The failure of Heavy Retraction to apply, despite the surface long vowel, justifies the intermediate step in the derivation. It is thus pointless to struggle to find ways to rule out DY derivations in OT, since such derivations actually do exist in human language.

3.4.4 *DY derivations in wider perspective*

Presumably, the reason why it is seen as important to rule out DY derivations in OT is to establish a difference in predictions between derivational phonology and OT. Ironically, the concept of "cumulativity of unfaithfulness," which was seen as an important step in ruling out DY derivations, actually increases the resemblance of OT to derivational theory, and thus strengthens the case that Sympathy Theory is a reconstruction of the intermediate step in derivational theory, not an independent concept. If a rule is applied to an underlying form as sketched in (120), this makes the form different from the underlying form, and results in IO faithfulness violations—thus a string is unfaithful to the input, that is, different from the input, only if a rule has applied.

(120)

	(via rules I, J, K)		(via more rules)	
/ABC/	→	WXY	→	[LMN]
		[differs from /ABC/ by changes $\alpha\beta\gamma$]		[inherits changes $\alpha\beta\gamma$, adds changes LMN]

A derivation from /ABC/ to intermediate *WXY* results from applying a set of rules {I,J,K}, which creates faithfulness violations $\{\alpha,\beta,\gamma\}$. A derivation which goes further necessarily first undergoes rules {I,J,K}, so the form must inherit the faithfulness violations of the intermediate form, and thus "cumulativity" is common to Sympathy and to derivational theory.

McCarthy (1999*d*) states that in DY derivations, "later steps do not accumulate the results of earlier steps, since some later step literally undoes the effect of an earlier step." Given that understanding of DY derivations, though, DY derivations have *never* been proposed in derivational theory and would run counter to the standard assumptions of derivational theory. In derivational theory, once a rule is applied, it cannot be "unapplied": time only flows forward. It is possible that some later rule can blindly assign a value which accidentally turns out to be the same as one found in an earlier stage. Even in derivational theory, a later rule never literally reaches back in time and undoes an earlier step—rules Markovianly apply only to what is locally available at the given moment.

(121)	/ba/	/bap/	/bat/	
	ba	ba	ba	deletion of final consonant
	[bat]	[bat]	[bat]	insertion of [t]: coincidentally the same *kind* of segment as was present underlyingly: *not* the same literal *token* of the consonant

The OT account of DY derivations also shares this essential property with derivational theory. A derivation of the form A → B → A is possible in OT, as long as "restored" A is token-wise distinct from underlying A.

The example from Zinza provides a model for handling one class of DY derivations in OT, namely those cases where an underlying object is deleted, and then another token of similar phonetic character is subsequently inserted. Not only will this work for tones and entire segments, but insofar as any featural change can be modeled as the deletion of one specification followed by the insertion of another, then it is probable that any delete-insert DY derivation can be handled.

There is also reason to believe that a DY derivation of the form insert–delete as proposed by Kiparsky, where /maat/→ *mač* via sympathy with *maači*, is actually possible, even in EST. Although the candidate *mač*, with no final vowel, will not survive the sympathy constraint because it lacks the epenthetic vowel, as can be seen in (122), there is a phonetically identical candidate (d), [mač<i>] with an epenthetic vowel which is phonetically unparsed. This tableau differs minimally from the one given by Kiparsky, only adding an explicit constraint to drive apocopation of the final vowel, which prohibits the last syllable from ending in a vowel.

(122)

		maat	*V]$_\sigma$#	❁Symp	Ident(hi)	Dep-V	*Max-μ
	a.	mat		*ℵ0!			*
❁	b.	maač	*		*	*	✓
	c.	mač		*ℵ0!	*		*
☞	d.	mač<i>			*	*	*

This last candidate thus inherits the Dep violation found in the flower candidate, but is otherwise phonetically identical to the candidate *mač* which was ruled out by its excessive faithfulness to the input. This indicates that DY derivations of the type insert–delete are not beyond the reach of OT, either. Whether such derivations are actually found in languages remains a matter for research; whatever the outcome of that search, there is no evidence that the ability to handle DY derivations distinguishes derivational phonology from OT.

3.5 Summary

In the course of this discussion, I have considered certain cases of serial derivation found in the Bantu languages of Tanzania, and have argued that a considerable range of theoretical devices proves necessary to handle these phenomena. It then remains a topic for future research to determine whether the devices that turned out to be required to handle the languages discussed here are sufficient to handle derivational concepts in non-Bantu languages, or in languages spoken outside of Tanzania, or whether further devices will need to be added to the theory.

4

Stress-Epenthesis Interactions

ELLEN BROSELOW

4.1 Introduction[1]

In many languages, stress assignment appears to ignore inserted vowels, giving rise to opaque stress patterns. This fact has supported arguments for multilevel derivations, which account for the apparent invisibility of epenthetic vowels by inserting them after stress is assigned. Because this approach requires multiple levels of derivation, stress-epenthesis interactions are potential problems for a framework that allows reference to only two levels, input and output. However, as Alderete (1995, 1999*b*) has argued, even strictly parallel versions of Optimality Theory can account for the invisibility of inserted vowels by means of constraints requiring elements in prosodic constituents to have correspondents in underlying representation. In this chapter I argue that the correspondence approach to stress-epenthesis interactions actually provides a better match with the wide array of stress-epenthesis interactions than the multi-level approach. The general argument of the chapter is that disruption of normal stress patterns by epenthetic material is caused by one of two factors: avoidance of epenthetic material in prominent positions, and maximization of paradigmatic contrasts. I discuss stress-epenthesis interactions in four languages. In Selayarese loanwords, the main stress foot is constructed to avoid inclusion of epenthetic vowels anywhere in the foot, while in North Kyungsang Korean loanwords and in Winnebago native vocabulary, epenthetic vowels are avoided in the head position of a foot. Iraqi Arabic illustrates a different motivation for the apparent invisibility of inserted vowels: the maximization of contrast between stems of different grammatical types.

[1] This study was supported by NSF grant SBR-9729108 to Daniel Finer and Ellen Broselow and by funding from NWO. I am greatly indebted to Hasan Basri for his penetrating insights into Selayarese structure; this chapter could not have been written without his work on Selayarese borrowings. I am also grateful to John Alderete, Lisa Selkirk, and Dan Finer for comments that have significantly improved the work, and to colleagues at Stony Brook and audiences at the Harvard/MIT P2K Workshop and the University of Leiden/HIL Colloquium Series for valuable discussion.

4.2 Selayarese loanwords

In Selayarese, one of the Makassar languages of South Sulawesi, Indonesia, the very general pattern of penultimate stress may be disrupted by the presence of epenthetic vowels. This disruption has been accounted for in a serial derivation by ordering epenthesis after stress assignment (Mithun and Basri 1986), and in strictly parallel OT by means of a constraint banning epenthetic material from the head foot of a prosodic word (Alderete 1995, 1999*b*). In terms of empirical coverage of native vocabulary, these two approaches are equivalent. However, data from loan phonology provides a wider range of stress-epenthesis interactions, and I argue below that only the strictly parallel approach can account successfully for the stress in borrowed words.

I begin by reviewing stress-epenthesis patterns in native vocabulary. Stress is normally penultimate in monomorphemic words, regardless of syllable structure:

(1) Normal penultimate stress

a.	sahála	'sea cucumber'
b.	palóla	'eggplant'
c.	balíkaʔ	'arm'
d.	sampúlo	'ten'
e.	búlaŋ	'moon, month'
f.	tímbo	'grow'
g.	góntiŋ	'scissors'
h.	barámbaŋ	'chest'
i.	kalihára	'ant'
j.	kalumánti	'big black ant'

This stress pattern can be analyzed as preference for a bisyllabic, trochaic foot at the right edge of the word. The bisyllabic nature of the foot is consistent with the minimal word size; all major category words consist of at least two syllables.

The exceptions to penultimate stress are of two kinds. First, several suffixal clitics fall outside the stress domain; these are argued by Selkirk (1999), Basri, Broselow, Finer (1999), and Basri, Broselow, Finer, and Selkirk (1997) to be outside the prosodic word. Second, there are a number of monomorphemic words with antepenultimate stress, which have been analyzed as containing a final epenthetic vowel (Mithun and Basri 1986; Piggott 1995; Basri, Broselow, Finer, and Selkirk 1997):

(2) Monomorphemes with antepenultimate stress

a.	sáhala	/sahal/	'profit'
b.	lámbere	/lamber/	'long'
c.	bótoro	/botor/	'gamble'
d.	sússulu	/sussul/	'burn'
e.	páʔrisi	/páʔris/	'painful'
f.	hállasa	/hallas/	'suffer'
g.	maŋkásara	/maŋkasar/	'Makassar'
h.	kasíssili	/kasissil/	'mosquito'
i.	barúasa	/baruas/	'cookie'
j.	salúara	/saluar/	'pants'

Comparison of (1a) *sahála* 'sea cucumber' and (2a) *sáhala* 'profit' reveals that stress cannot be entirely predicted from surface structure. However, all morphemes with antepenultimate stress share certain properties. First, all end in a vowel which is preceded by /*r*/, /*l*/, or /*s*/, none of which is an acceptable coda in this language. The vowel following /*r*, *l*, *s*/ is identical to the preceding vowel, suggesting that a copy of the nearest vowel is inserted to allow stem-final /*r*, *l*, *s*/ to be syllabified as an onset. This analysis is confirmed by the fact that final vowels of stems with antepenultimate stress disappear before a vowel-initial suffix (3a,b), in contrast with other final vowels, as in (3c,d):

(3) Disappearance of epenthetic vowel

a.	lámbere	lambéraŋ	/lamber+aŋ/	'long/longer'
b.	hállasa	hallási	/hallas+i/	'suffer/make suffer'
c.	tirére	tireréaŋ	/tirere+aŋ/	'thirsty/thirstier'
d.	rúppa	ruppái	/ruppa+i/	'face/confront'

In a serial approach, stress is assigned to the penultimate syllable before the final vowel is inserted. In a parallel approach, the constraint Head-Dep (Alderete 1995, 1999*a*) prevents the main stress foot from containing epenthetic material. This means that in a word like *lámbere*, from underlying */lamber/*, the bisyllabic trochaic stress foot is built on the first two vowels, leaving the final vowel unfooted: {*lámbe}rE*. (Here and in the following discussion, inserted vowels are shown in upper case.)

These two analyses do equally well in accounting for forms with final epenthetic vowels. The more interesting cases, however, involve medial epenthesis, which we can see in the adaptation of loanwords that do not conform to Selayarese phonotactic constraints. Most loans into Selayarese are from Bahasa Indonesia (BI), the lingua franca of the region. In general, the stress of the BI forms is ignored, with loans stressed on their penultimate syllable. However, as the forms below illustrate, BI forms with final /*r*, *l*, *s*/ undergo epenthesis and are stressed on their antepenultimate syllables, just like native vocabulary (Basri 1997).[2]

(4) Loans with final epenthesis: σ'σE

a.	bótolO	'bottle'	BI: bótol
b.	sénterE	'flashlight'	BI: sénter
c.	kálasA	'class'	BI: kəlás
d.	bérasA	'rice'	BI: bərás
e.	kábalA	'cable'	BI: kábal
f.	kábarA	'news'	BI: kábar
g.	kíkirI	'metal file'	BI: kíkir

This confirms the invisibility of final epenthetic vowels for the purposes of stress. In contrast, however, epenthetic vowels in penultimate position must be visible—they themselves bear stress:

(5) Loans with medial epenthesis: σE'σ

karÁtu	'card'	BI: kártu
surÚga	'heaven'	BI: súrga
bakÁri	proper name	BI: bákri
burÚhaŋ	proper name	BI: búrhan
ramÁli	proper name	BI: rámli

Quadrisyllabic words further complicate matters. As in native vocabulary, stress is penultimate when the two final syllables are underlying, and antepenultimate in forms with a final epenthetic vowel:

(6) Quadrisyllabic Loans:[3]

a. σEσ'σ, Eσσ'σ

samAsúddiŋ	proper name	BI: syamsúddin
pArajúriʔ	'soldier'	BI: prajúrit

b. σ'σσE

balábasA	'ruler'	BI: bələbás

However, when both the last and the third-from-last vowel are epenthetic, stress falls on the penultimate syllable:

[2] See Broselow (2000) for discussion of forms ending in consonants other than /*r*,*l*,*s*/.

[3] The symbol 'y' indicates a palatal glide and 'j' a voiced palatal stop.

(7) Quadrisyllabic Loans: σEσ’E

solOdérE	‘weld’	BI: sólder
korOnélE	‘corner kick (in soccer)’	BI: kórnel
karAtísI	‘ticket’	BI: kárcis
tarApálA	‘tarpaulin’	BI: térpal
tapAsérE	‘interpretation’	BI: tápsir

The generalization, then, is that only in final position does an epenthetic vowel disrupt the normal penultimate stress pattern, yielding antepenultimate stress. However, when the final epenthetic vowel is accompanied by another epenthetic vowel two syllables to its left, the final vowel must count in the stress computation.

This pattern is problematic for the serial analysis. We can easily derive the discrepancy between final and medial epenthesis in trisyllabic forms by assuming that word-final consonants are extraprosodic, and not syllabified until late in the derivation. Medial /*r*,*l*,*s*/ will be syllabified before stress is assigned, while final /*r*,*l*,*s*/ may remain in limbo until some later point in the derivation, as illustrated in (8):

(8) Serial analysis:

	a. /sahal/	b. /kartu/
Final extrametricality:	saha (l)	—
Syllabification, epenthesis:	sa.ha (l)	ka.rA.tu
Stress assignment:	{sá.ha}(l)	ka.{rÁ.tu}
Loss of extrametricality; Syllabification, epenthesis:	{sá.ha.} lA	ka.{rÁ.tu}

The invisibility of epenthetic vowels following stem-final consonants then follows from the invisibility of stem-final consonants. However, this approach predicts the wrong output for forms with both final and medial epenthesis:

(9)

	a. /solder/	b. /balabas/
Final extrametricality:	solde(r)	balaba(s)
Syllabification, epenthesis:	so.lO.de (r)	
Stress assignment:	so.{lÓ.de} (r)	ba{lába}(s)
Loss of extrametricality; Syllabification, epenthesis:	*so.{lÓ.de.}rE (solO{dé.rE})	ba{lába}sA

Form (9a) receives antepenultimate stress, rather than the actual penultimate stress. To derive the correct stress, we would need a stress-readjustment rule converting the antepenultimate stress of (9a) to the correct penultimate stress. But such a rule would need to leave intact stress on forms like (9b). Since the metrical structure of (9a) and (9b) is equivalent at the point this rule would apply, the rule would need to be non-Markovian, distinguishing underlying from inserted vowels.[4]

This distinction is of course at the heart of the strictly parallel approach. In this approach, stress feet are constructed (where possible) on underlying vowels only. The inclusion of epenthetic vowels in the stress foot (in either head or non-head position, as in *ka{rÁtu}* 'card' or *solO{dérE}* 'weld') occurs only when it is impossible to construct a bisyllabic foot that does not contain an epenthetic vowel. The following constraints derive this pattern:

(10) Selayarese Stress Constraints

a. FT BIN(σ), FT TROC: Feet are bisyllabic and trochaic. These constraints are ranked so high as never to be violated.

b. HEAD-DEP (Alderete 1999*a*,*b*,*c*): Every vowel contained in a prosodic head in S_2 has a correspondent in S_1 (i.e., vowels in prominent foot must not be epenthetic).[5]

c. ALIGN-R (PWD, FT): The right edge of the prosodic word should be aligned with the right edge of a foot.

Constraints (10a,c) enforce the normal penultimate stress pattern. Ranking HEAD-DEP over ALIGN-R will choose antepenultimate stress for trisyllables with final epenthesis, where the stress foot includes only lexical vowels. But for forms with medial epenthesis, there is no possible parse into bisyllabic feet, and therefore the best that can be done is to satisfy the requirement that the foot be right-aligned, yielding penultimate stress, as in (13). The loanword data therefore provide striking confirmation of Alderete's analysis of native vocabulary:

(11) /sahala/ 'sea cucumber' σσ'σ		FTBIN, FTTROCH	HEAD DEP	ALIGN-R (PWD,FT)
☞	a. sa {hála}			
	b. {sáha}la			*!

[4] A reviewer asks whether in fact penultimate stress is simply the default for quadrisyllabic forms, with *balábasa* representing an exceptional pattern. As the native forms (2g,h,i,j) illustrate, the antepenultimate stress in *balábasa* is typical of forms with final epenthetic vowels.

[5] Alderete's proposed constraint is more (probably too) general, banning any epenthetic material from the prosodic head. Nothing hinges on this distinction here.

(12) /sahal/ 'profit' σ'σE			
a. sa {hálA}		*!	
☞ b. {sáha}lA			*
(13) /kartu/ 'card' σE'σ			
☞ a. ka {rÁtu}		*	
b. {kárA}tu		*	*!

In quadrisyllabic forms, the principle is the same–the ideal parse constructs a bisyllabic trochaic foot aligned with the right edge of the word, but if the final vowel is epenthetic, the alignment requirement is overridden, moving the stress foot one syllable leftward. (Below I show only parses containing a single stress foot; for a full treatment of possible outputs, see Broselow (2000)).

(14) /kalihara/ 'ant' σσσ'σ	FtBin, Troch	Head-Dep	Align-R (PWd, Ft)
a. {káli} hara			*!*
b. ka {líha} ra			*!
☞ c. kali {hára}			
(15) /maŋkasar/ 'Makassar' σσ'σE			
a. {máŋka} sarA			**!
☞ b. maŋ {kása} rA			*
c. {maŋka} {sárA}		*!	

Where both the final and the antepenultimate vowels are epenthetic, it is impossible to construct a bisyllabic foot that does not contain an epenthetic vowel, which makes Head-Dep irrelevant and the alignment constraint decisive:

(16) /solder/ 'weld' σEσ'E	FTBIN, TROCH	HEAD-DEP	ALIGN-R (PWD, FT)
b. {sólO}derE		*	*!*
c. so {lÓde} rE		*	*!
☞ d. solO {dérE}		*	

Thus, the strictly parallel analysis predicts that an epenthetic vowel disrupts the normal construction of a bisyllabic foot aligned with the right edge only when it is possible, by shifting the foot over, to construct a foot containing only underlying vowels. The serial account, on the other hand, provides no account for why a final epenthetic vowel should be invisible when preceded by two underlying vowels, but visible when preceded by an antepenultimate epenthetic vowel.

We might attempt to save the serial analysis by employing the Domino Condition (Halle and Vergnaud 1987), which directs that when material is inserted into a foot, that foot and all feet to its right/left are destroyed (moving toward the edge from which feet are constructed, or with which feet are aligned). Stress is then reassigned only on the liberated portions of the word, including the inserted material in the computation. In Selayarese, this means that epenthetic vowels to the right of the penultimate underlying vowel should cause a reversion to default stress. Hayes (1995) points out empirical problems with the Domino Condition—in some cases, it simply makes the wrong predictions. For Selayarese, however, an analysis using the Domino Condition suffers from conceptual problems. First, assuming feet are constructed on syllables, this analysis would require us to allow /*r*, *l*, *s*/ in forms like /*solder*/ to be syllabified in coda at some level, only to trigger epenthesis at some later level. But it is unclear why epenthesis is motivated at all, if these consonants can be syllabified in the coda (and we cannot appeal to extrametricality without giving up the generalization that only segments at edges are extrametrical). Even ignoring these problems, however, the Domino Condition is less satisfying than the parallel account in that it simply stipulates the connection between the direction of foot destruction and foot construction. This stipulation (that if feet are constructed from the right (left), then epenthesis into a foot entails destruction of all feet to the right (left) of the invaded foot) is a way of ensuring that default stress arises when an epenthetic vowel occurs in a main stress foot position. In the parallel analysis, this generalization falls out of the ranking HEAD-DEP≫ ALIGN-R: the reversion to default penultimate stress occurs when violations of HEAD-DEP cannot be avoided. Thus, while the serial analysis of Selayarese might be salvaged by adding the Domino Condition to the rules of the grammar, the parallel analysis derives the same generalizations from a set of ranked constraints. In the next section, we will see another case of stress-epenthesis interactions for which the Domino Condition makes the wrong predictions.[6]

4.3 North Kyungsang Korean loanwords

A second example of stress-epenthesis interaction in loanwords is provided by borrowings into North Kyungsang Korean discussed by Kenstowicz and Sohn (2000). Kenstowicz and Sohn report that this dialect of Korean (henceforth, NKS Korean) is characterized by a pitch accent system in which each word must have at least one pitch peak. There are some subregularities in the pitch accent system: words with a long vowel in the first syllable generally have an HH pattern, and words longer than three syllables most often have penultimate accent. But to a large extent, the native language accent pattern is lexically determined, as illustrated by the following contrasts:

(17) North Kyungsang Native Accent
  a. HH
    hárépi 'grandfather'
  b. HL
    kámani 'rice bag'
    káci 'kind'
  c. LH
    kurúma 'cart'
    kací 'eggplant'

Like longer native forms, loans generally have penultimate accent (18a), though there is some evidence of a preference for accenting a final heavy over a penultimate light syllable (18b). It is facts like these that lead Kenstowicz and Sohn (2000) to argue that NKS Korean accent in loans provides an example of emergence of the unmarked, in the form of a preference for a Romance-type metrical structure:

[6] Another possible argument for the serial analysis of epenthesis is provided by Piggott (1995), who notes, following Mithun and Basri (1986), that while underlying stressed vowels in open syllables are lengthened, presumably to satisfy a bimoraic minimum requirement, epenthetic vowels fail to lengthen. However, Basri (1999) provides an analysis of these facts in a parallel framework. See the Appendix for fuller discussion of this issue.

(18) Loan Accent

a. penultimate accent

k^{h}ít^{h}a	‘guitar’
amerík^{h}a	‘America’
k^{h}ellip^{h}onía	‘California’

b. final accent

k^{h}epinét	‘cabinet’

The accent patterns illustrated in (18) can be accounted for by assuming a preference for bimoraic trochaic feet.

NKS Korean borrows freely from English, and many borrowings undergo epenthesis. While NKS borrowings, like Selayarese borrowings, attest to a preference for penultimate stress as the default pattern, the borrowing languages contrast with respect to the behavior of forms with epenthetic vowels. In NKS Korean loanwords, final epenthetic vowels appear to be visible, in contrast with such vowels in Selayarese:

(19) σσE, EσE

t^{h}enísU	‘tennis’
te.í.t^{h}U	‘date’
ma.ú.sU	‘mouse’
ma.í.k^{h}U	‘mike’
kUrásU	‘glass’
kUllápU	‘glove’

However, epenthetic vowels in penultimate position, which take the stress in Selayarese, are generally not accented in NKS Korean. When the two final vowels are epenthetic, accent falls on the antepenult, while a word with a single epenthetic vowel in penultimate position takes accent on its final syllable:[7]

(20) a. σ’EE

t^{h}ósUt^{h}U	‘toast’
pésUt^{h}u	‘best’
réphUt^{h}U	‘left’
t^{h}éksUt^{h}U	‘text’
kíp^{h}Ut^{h}U	‘gift’
p^{h}ásUt^{h}U	‘first’

[7] Kenstowicz and Sohn (2000) note that some forms (pakÚna ‘Wagner’, rarÚko ‘largo’) do have accent on a penultimate epenthetic syllable; they speculate that these are older forms in which the inserted vowel has been reinterpreted as underlying.

b. σEσ'
   met^{h}Uró 'metro'
   nigUró 'negro'
   k^{h}ont^{h}Uról 'control'

Thus, while both Selayarese and NKS Korean exhibit disruption of the generally preferred penultimate stress pattern in the presence of epenthetic vowels, the disruptions are of a different type. In Selayarese, disruption is associated with final but not penultimate epenthetic vowels, while in NKS Korean it is penultimate epenthesis that is disruptive. We can account for these differences by assuming that while Selayarese avoids incorporating an epenthetic vowel into any position in the main stress foot, NKS Korean simply avoids allowing an epenthetic vowel in the prominent (accented) position. Thus, while HEAD-DEP ranks relatively low in NKS Korean, the following constraint is highly ranked:

(21) HEADSYLL-DEP (Alderete 1995): Every segment contained in the head of a foot in S_2 has a correspondent in S_1 (epenthetic vowels cannot be the head of a foot).

In Selayarese, the ranking HEAD-DEP >> ALIGN-R accounts for the leftward shift of stress in forms with final epenthesis. In NKS Korean, ALIGN-R dominates HEAD-DEP, giving penultimate accent so long as this accent does not fall on an epenthetic vowel:

(22) /t^{h}enis/	HEADSYLL-DEP	ALIGN-R (PWD, FT)	FTBIN	HEAD-DEP
☞ a. t^{h}e{nísU}				*
b.{t^{h}éni} sU		*!		
c. t^{h}eni {sÚ}	*!		*	*
d. t^{h}e {ní}sU		*!	*	
e. {t^{h}é}ni sU		*!*	*	

However, accent does shift leftward when both final and penultimate vowels are epenthetic, due to high-ranking HEADSYLL-DEP:

(23) /t^{h}ost^{h}/	HEADSYLL-DEP	ALIGN-R (PWD, FT)	FTBIN	HEAD-DEP
a. t^{h}o{sÚt^{h}U}	*!			*
☞ b. {t^{h}ósU}t^{h}U		*		*
c. t^{h}osU{t^{h}Ú}	*!		*	*
d. t^{h}o{sÚ}t^{h}U	*!	*	*	*
e. {t^{h}ó}sUt^{h}U		**!	*	

When only the medial vowel is epenthetic, the best parse is a (non-binary) right-aligned foot:

(24) /met^{h}tro/	HEADSYLL-DEP	ALIGN-R (PWD,FT)	FTBIN	HEAD-DEP
a. me{t^{h}Úro}	*!			*
b. {mét^{h}U}ro		*!		*
☞ c. met^{h}U{ró}[8]			*	
d. me{t^{h}Ú}ro	*!	*	*	*
e. {mé}t^{h}Uro		**!	*	

It is difficult to see how these facts could be accounted for in a serial framework. Forms like *k^{h}ít^{h}a* 'guitar' and *ameríkh$^{}$a* 'America' illustrate that penultimate accent is preferred, motivating a highly ranked constraint demanding trochaic feet. Forms like *t^{h}enísU* 'tennis' indicate that epenthetic vowels, even final ones, should be present when stress is assigned. But if that is the case, we would expect penultimate stress in forms like *t^{h}ósUt^{h}U* 'toast' and *met^{h}Uró* 'metro' (**t^{h}osÚtU*, **met^{h}Úro*). Note that the Domino Condition is useless here, since that condition predicts that insertion of material to the right of the penultimate underlying vowel should cause destruction of foot structure. Reassignment of accent to the liberated material, including the inserted vowel(s), should then yield default penultimate accent. But in NKS Korean,

[8] A reviewer points out that the same metrical pattern could be arrived at with a different metrical parse, *me{t^{h}Uró}*, in which the two final syllables are grouped into an iambic foot. Under this parse, HeadSyll-Dep would force a violation of TROCHFT, rather than a violation of FTBIN.

insertion of a vowel into an existing foot ({*met^{h}ro*}, {*t^{h}ost^h*}) yields either antepenultimate or final accent.

In contrast, the parallel approach (with correspondence constraints) not only accounts for the data, but also provides insight into the similarities and differences between NKS Korean and Selayarese loanword adaptation. Selayarese and NKS Korean are alike in avoiding epenthetic vowels in prosodically prominent positions, even at the cost of sacrificing alignment of the main foot with the right edge of the word. They differ, however, in their definitions of prominent position (anywhere in the head foot vs. in the prominent position in a foot). They differ as well in the relative rankings of FtBin and Align-R; Selayarese is unyielding in its requirement that feet be bisyllabic, while NKS Korean is willing to sacrifice binarity for the sake of right-alignment.

Another respect in which the two languages differ is the extent to which the rankings of the relevant constraints are motivated by the native vocabulary. In Selayarese, the native vocabulary, while providing evidence for epenthesis in a much smaller range of cases, still motivates the rankings necessary to handle the loanword data. NKS Korean native vocabulary, in contrast, provides no obvious evidence for high ranking of HeadSyll-Dep, suggesting that this may be an instance of the emergence of the unmarked. We now turn to another case illustrating the role of HeadSyll-Dep, this time in native vocabulary.

4.4 Winnebago

The problem of stress-epenthesis interactions in the Siouan language Winnebago has received a great deal of attention (e.g. Miner 1979, 1981, 1989; Hale and White Eagle 1980; Hale 1985; Halle and Vergnaud 1987; Steriade 1990; Hayes 1995; Halle and Idsardi 1995; Alderete 1995). Because a number of researchers have provided analyses in a serial framework, it is important to determine whether these facts can be accounted for in a framework with only two levels. I argue that Winnebago, like NKS Korean, illustrates avoidance of foot heads containing epenthetic nuclei. Authorities differ on whether Winnebago should be considered to employ a stress system or a pitch accent system; I will assume that Winnebago employs a system of accent, with the position of the accent determined by metrical foot structure.

The facts of Winnebago accent are complex. Below I indicate only primary accent. (Nasalization, which is irrelevant to the analysis, is also not indicated.) In words with only light syllables, the accent falls on the third syllable and every other syllable thereafter (except in bisyllables, which have accent on their rightmost syllable). In forms beginning with a heavy syllable, accent falls on the second syllable, and on subsequent even-numbered syllables:

(25) a. All light

wadʒé	'dress'
hotaxí	'expose to smoke'
haratʃábra	'the taste'
hokiwároké	'swing (n.)'

b. Initial heavy

maatátʃ	'promise (1sg.)'
waakítʔe	'speak to (1sg.)'
waipéresgá	'linen'

I will assume, following Miner (1979, 1981, 1989) and Hayes (1995), that syllables are grouped into iambic feet, with accent falling on each syllable following a foot. The following constraints derive the patterns in (26):[9]

(26) a. FtBin (Mora), Ft=Iambic

b. Align-L (PWd, Foot): Align left edge of Prosodic Word with left edge of a foot.

c. PostAccenting: The syllable to the right of a foot should be accented.

d. *Accent: vowels should not be accented (no accented vowels unless required to satisfy constraints).

4.4.1 *Accent and Epenthesis*

Winnebago has an epenthesis process known as Dorsey's Law by which a vowel is inserted between a voiceless obstruent and a following sonorant consonant. The inserted vowel is a copy of the following vowel. These inserted vowels may be associated with disruption of the normal accent patterns, as illustrated by comparison of quadrisyllabic forms with and without inserted vowels:

(27) LLLL words

a. no epenthesis: {σσ}{σ'σ}

haratʃábra 'the taste'

b. normal accent: Eσσ'σ, σσE'σ, EσE'σ

kEredʒúsep	'Black Hawk'
hanipʃÁna	'I swam (declar.)'
kEreʃkÉreʃ	'colorful'

[9] John Alderete points out (pers. comm.) that this analysis of Winnebago avoids the necessity for positing initial extrametricality (otherwise quite rare).

c. disrupted accent: σEσσ’
hikOrohó ‘prepare, dress (3sg.)’

As the forms above illustrate, an epenthetic vowel disrupts accent when it occurs in the second syllable from the left (though only in words longer than three syllables). In contrast, epenthetic vowels in odd-numbered syllables are associated with normal accent. These patterns can be accounted for by assuming that HeadSyll-Dep plays a leading role in Winnebago, as in the adaptation of NKS Korean loanwords. Normally, an iambic foot is formed at the left edge of the word, with accent falling on the syllable following this foot (that is, the third syllable). Thus, alignment of the foot with the left word edge places an accent on the third syllable. But the normal accent pattern is disrupted just when the syllable that should be the head of a foot is epenthetic. In this case, the ranking HeadSyll-Dep >> Align-L will shift the iambic foot one syllable to the right, choosing the parse *hi{kOro}hó* ‘prepare, dress (3sg.)’ over the well-aligned **{hikO}{róho}*, for example. The following tableaux illustrate the array of LLLL word types:

(28) LLLL, no epenthesis /haratʃabra/ σσσ’σ ‘the taste’	FtBin, FtIamb	HeadSyll-Dep	Align-L (PWd, Ft)	Post Accent
☞ a. {hara}{tʃábra}				
b. ha{ratʃa}brá			*!	
c. {ha}{rátʃa}{brá}	*!*			

(29) LLLL, normal accent /kreʃ kreʃ/ EσE’σ ‘colorful’	FtBin, FtIamb	HeadSyll-Dep	Align-L (PWd, Ft)	Post Accent
☞ a. {kEreʃ}{kÉreʃ}				
b. kE{reʃ kE}réʃ		*!	*	

(30) LLLL, disrupted accent /hikroho/ σEσσ’ ‘prepare’	FtBin, FtIamb	HeadSyll-Dep	Align-L (PWd, Ft)	Post Accent
a. {hikO}{róho}		*!		
☞ b. hi{kOro}hó			*	

We next consider trisyllabic words. As illustrated below, accent is never disrupted in trisyllabic forms, even when the epenthetic vowel is the second in the word:

(31) LLL words

a. no epenthesis: {σσ}σ'
   hotaxí 'expose to smoke'

b. normal accent: Eσσ', σEσ'
   ʃUruʃgé 'you (sg.) untie it'
   hokEwé 'enter'

I have argued that in Winnebago, accent falls on the syllable following each foot. However, footing the first two syllables of *hokEwé* would place the epenthetic vowel in head position. However, the universal constraint set must contain a constraint HEADSYLLACCENT which produces the familiar pattern of accent on the head syllable. In Winnebago, this constraint is normally masked by higher ranked POSTACCENT, which assigns accent to the syllable following the foot, and OCP, which forbids retention of adjacent accents. But the effect of HEADSYLLACCENT emerges in forms with an epenthetic second syllable; because there is no syllable following this foot to receive the accent, HEADSYLLACCENT can be satisfied. Thus, trisyllabic forms will receive accent on their final syllable either by accent on the post-foot syllable, or accent on the head syllable of the foot:

(32) Stress/Accent Constraints (Final Version)

a. FTBIN (MORA), FT=IAMBIC

b. HEADSYLL-DEP: Every segment in the head of a foot in S_2 has a correspondent in S_1.

c. ALIGN-L (PWD, FOOT): Align left edge of Prosodic Word with left edge of a foot.

d. OCP(ACCENT): Adjacent syllables may not be accented.

e. POSTACCENTING: The syllable to the right of a foot should be accented.

f. HEADSYLLACCENT: The head of a foot should be accented.[10]

[10] We will also need a constraint *ACCENT (vowels should not be accented), which prevents accents from surfacing at random.

(33) /hotaxi/ 'expose to smoke' σσσ'	FtBin, FtIamb	HeadSyll-Dep	Align-L	OCP	PostAccent	HeadSyllAccent
☞ a. {hota}xí						*
b. ho{taxí}			*!			
(34) /ʃruʃge/ 'you (sg) untie it' Eσσ'	FtBin, FtIamb	HeadSyll-Dep	Align-L	OCP	PostAccent	HeadSyllAccent
☞ a. {ʃUruʃ}gé						*
b. ʃU{ruʃgé}			*!			
(35) /hokwe/ 'enter' σEσ'	FtBin, FtIamb	HeadSyll-Dep	Align-L	OCP	PostAccent	HeadSyllAccent
a. {hokE}wé		*!				*
☞ b. ho{kEwé}			*			

These constraints account equally well for words of more than four syllables. Normally, accent falls on the third and following odd-numbered syllables, but again, accent is disrupted by an epenthetic vowel in an even-numbered syllable:

(36) Longer words
  a. no epenthesis
    hokiwároké — 'swing (n.)'
  b. normal accent
    hirakÓrohó — 'prepare, dress (2sg.)'
    hirakÓrohónirá — 'the fact that you do not dress'
  c. disrupted accent
    wakIripÁras — 'flat bug'
    wakIripÓropÓro — 'spherical bug'
    harakíʃUrudʒíkʃAná — 'pull taut (2sg. declar.)'

Longer forms containing odd-numbered syllables could conceivably be parsed in different ways; for example, the accent pattern of (37) is consistent with footing (37a), in which there is a final stray syllable, or (37b), with a medial stray syllable. The constraint ranking proposed here chooses (37b), since this satisfies both PostAccent and HeadSyllAccent. The same footing is possible for (38), since it does not require creating a foot which has an epenthetic vowel as its righthand (head) element. However, forms like (39), in which the second syllable is epenthetic, require the shifting of feet to the right:

(37) no epenthesis /hokiwaroke/ 'swing (n.)' σσσ'σσ'	FtBin, FtIamb, Parse-2	HeadSyll Dep	Align-L	OCP	Post Accent	HeadSyll Accent
a. {hoki}{wáro}ké						**!
☞ b. {hoki}wá{roké}						*
c. ho{kiwa}{róke}			*!			**
(38) normal accent /hirakroho/ 'prepare' σσE'σσ'						
a. {hira}{kÓro}hó						**!
☞ b.{hira} kÓ{rohó}						*
c. hi {rakO}{róho}		*!	*			**
(39) disrupted accent /wakripras/ 'flat bug' σEσE'σ						
a. {wakI}{rípA}rás		**!				**
b. {wakI} rí {pÁrás}		*!				*
☞ c. wa {kIri}{pÁras}			*			**

The constraint set developed for light syllables will account equally well for accent-epenthesis interactions in words containing heavy syllables. Accent falls on a syllable following an initial heavy syllable, whether that syllable is underlying or epenthetic. Thus, in light-syllabled forms, an epenthetic vowel in the second position disrupts the normal accent pattern (compare *haratʃábra* 'the taste' with no epenthesis and *hikOrohó* 'prepare, dress (3sg)' with second vowel epenthetic). In contrast, forms like (40a) and (40b) have the same accent pattern:

(40) a. {haa}kí{tujík} ‘I pull it taut (plain)’

b. {waa}pÓ{rohí} ‘snowball making’

c. {waa}{pÓro}{pÓro} ‘snowball’

This follows if the initial heavy syllable itself constitutes a foot, which then causes the following syllable to be postaccented.[11]

4.4.2 *Previous Analyses of Epenthesis-Accent Interactions*

In the analysis proposed above, the disruption of normal accent by epenthesis in an even-numbered syllable stems from the high rank of HEADSYLL-DEP, which disallows feet of the form {σE}. Thus, although the normal footing is {σσ}{σσ} (as in *{hara}{tʃábra}* ‘the taste’), the sequence σEσσ will be footed as σ{Eσ}σ (as in *hi{kOro}hó* ‘prepare’), because an epenthetic vowel cannot be the head (rightmost) syllable of a foot.[12]

Alternative analyses derive the impossibility of creating a foot of the form {σE} in different ways. The analysis of Halle and Idsardi (1995) posits a constraint requiring an epenthetic syllable to coincide with a left metrical constituent boundary. Like HEADSYLL-DEP, this constraint rules out feet of the form {σE}. But while the HEADSYLL-DEP analysis relates this prohibition to universal constraints against epenthetic material in prominent positions, the analysis using a left-coincidence constraint does not.

Other analyses rely on the assignment of somewhat unorthodox syllable structures to σE sequences. Thus, to prevent the formation of {σE} feet, Hayes (1995) argues that there is a level prior to footing at which a sequence like */kro/* in */hikroho/* ⇒ *hikOrohó* would constitute a single syllable. To account for the position of accent on the syllable following this sequence, he assumes further that a sequence like */kro/* constitutes a heavy syllable, with both the vocalic nucleus and the onset sonorant consonant bearing a mora. These sequences then pattern with other heavy syllables in taking accent on the syllable following them. However, since a structure like *kOro* patterns with two light syllables with respect to his tone-shift rule, he must assume that this sequence is transformed into two light syllables by the point at which tone shift applies. Similarly, Alderete (1995), though working within a strictly parallel framework, argues that a sequence like *kOro* constitutes a single heavy syllable in the output, though it is presumably realized phonetically as two syllables. Note that this approach cannot be extended to Selayarese,

[11] Forms with non-initial heavy syllables are discussed in the Appendix.

[12] The constraint LAPSE-2, which prevents a sequence of two unfooted syllables (Alderete 1999*b*), will rule out hikO{roho}.

in which treating *solOdérE* as a bisyllabic form would have disastrous results.

All these analyses are designed to force an epenthetic vowel to form a foot with a following rather than a preceding vowel. In the analysis proposed here, this follows from high ranking of HEADSYLL-DEP, which bans epenthetic vowels from the right (head) syllable of a foot.

4.5 Iraqi Arabic

I now turn to a disruption of normal stress by an inserted vowel that does not lend itself to the sort of account outlined above. I will suggest that this stress disruption is due to morphological factors rather than to the presence of epenthetic vowels.

The relevant fact is the apparent invisibility of epenthetic vowels to stress assignment in Iraqi Arabic. Stress is quite regular in this dialect. As shown below, stress falls on the final syllable if that syllable consists of a long vowel followed by a consonant; on the penultimate syllable if the penultimate is heavy (containing a long vowel or closed by a consonant); and otherwise on the antepenultimate syllable:

(41) Iraqi Arabic stress:
  a. final syllable: kitáab 'book'
  b. heavy penult: sallátha 'her basket', ʕiráaqi 'Iraqi'
  c. antepenult: ʃárika 'company', ʕáalami 'world', mumáθθ ila 'actress'

This regular pattern is disrupted, however, in the presence of epenthetic vowels: in *kitábIt* 'I wrote / you (2 sg. m.) wrote' the suffix consists of /t/, but a vowel is inserted to prevent a complex coda. (This dialect allows only one consonant in coda and in onset, except in word-initial position, where complex onsets are possible). Stress falls on a light penultimate syllable, in contrast to *ʃárika*, which has identical surface syllable structure and the expected antepenultimate stress.

Forms like *kitábit* 'I wrote / you (m. sg.) wrote' are actually anomalous in another respect as well. Comparison of the full perfect tense paradigm reveals that the final vowel of a CVCVC verb stem is normally deleted when a vowel follows the stem, as in the third-person singular feminine and third-person plural forms:

(42) 'write (perfect)'

kítab	'3 sg. m.'	/kitab/
kítbat	'3 sg. f.'	/kitab+at/
kítbaw	'3 pl.'	/kitab+aw/
kitábit	'2 sg. m.'	/kitab+t/
kitábti	'2 sg. f.'	/kitab+ti/
kitábtu	'2 pl.'	/kitab+tu/
kitábit	'1 sg.'	/kitab+t/
kitábna	'1 pl.'	/kitab+na/

Thus, the *kitábit* forms are opaque with respect to both stress and syncope. In a serial analysis, this opacity can be accounted for by ordering stress and syncope rules before epenthesis:

(43) Serial analysis (Broselow 1982):

	a. /kitab+at/	b. /kitab+t/
syncope:	kitbat	—
syllabification:	kit.bat.	ki.tab.t
stress assignment:	kít.bat	ki.táb.t
epenthesis:	—	ki.táb.It
resyll:	—	ki.tá.bIt
	[kítbat]	[kitábit]
	'she wrote'	'I wrote/ you (sg. m.) wrote'

In this analysis, the stem actually contains a heavy penult at the point at which stress is assigned. The anomalous stress in the first-person singular and the second-person masculine singular follows from the fact that these forms are the only ones to take a suffix consisting of a single consonant.

It is more difficult to see how the stress disruption in these forms could be treated within a strictly parallel framework. Note that this pattern is crucially different from those we have considered earlier. In Selayarese, NKS Korean, and Winnebago, when the normal patterns of foot construction would place epenthetic material in prominent positions, feet were shifted to include only underlying vowels. But in this case, the expected stress pattern *{*kíta*}*bIt* would involve a foot that does not contain any epenthetic material (according to the normal assumption that stress in this language involves a bisyllabic trochee). Therefore, neither Head-Dep nor HeadSyll-Dep should prevent assignment of the foot structure found in *ʃárika* to the form *kitábIt*. These facts, therefore, seem to favor the serial analysis.

However, while the serial account is appealing, this account does not extend to other verb types, in which we see anomalies that do not receive an intuitively satisfying phonological explanation. I will argue, therefore, that the stress disruption in forms like *kitábIt* is due not to the presence of an epenthetic vowel but instead to a more general phenomenon. In this dialect (as in many of the colloquial Arabic dialects), the base of suffixation in third-person perfective verb forms is always distinct from first- and second-person verb bases.[13] A survey of different verb-stem shapes in the perfective is instructive. We begin with triconsonantal verbs, which in their unsuffixed form are bisyllabic:

(44) Triconsonantal Verbs

a. 'write'
3rd kítab, kítbat, kítbaw
2nd kitábit, kitábti, kitábtu
1st kitábit, kitábna

b. 'telephone'
xáabar, xáabrat, xáabraw
xaabárit, xaabárti, xaabártu
xaabárit, xaabárna

c. 'change'
báddal, bádlat, bádlaw
baddálit, baddálti, baddáltu
baddálit, baddálna

The third-person forms all have stress on the initial stem syllable, while the others have stress on their second syllable. The phonological analysis of these facts derives these differences from the suffix shape: [+3p] suffixes are zero, or vowel-initial *(+at, aw)*, while [−3p] suffixes are (at least underlyingly) consonant-initial (*+t, +ti, +tu, +t, +na*).

The phonological analysis is no doubt a good explanation of how the differences between these verb bases arose, but does not necessarily provide the best account of the synchronic facts. Consider the so-called final weak verbs, where we see differences between [+3p] and [−3p] bases which go beyond the stress:

(45) Final Weak Verbs

	'forget'	
3rd	nísa, nísat, nísaw	(+∅, +at, +aw)
2nd	niséet, niséeti, niséetu	(+t, +ti, +tu)
1st	niséet, niséena	(+t, +na)

[13] Many years ago, Bob Harms suggested a similar, functionally based analysis in a class at the University of Texas at Austin. It should be noted that not all dialects impose a distinction in the shape of [+3] and [−3] perfective verb bases; for example, a Bedouin dialect of the Cyrenaican Jebel discussed by Mitchell (1960) has *kitáb* 'he wrote' and *kitábit* 'I/you m. wrote'. In this dialect, however, stress always falls on the final syllable of the perfective base, whether the suffix contains an underlying vowel or an inserted vowel.

These verbs historically have a glide as their final radical, and Brame (1970) has argued for a synchronic analysis in which the final glide is still present. The glide is deleted word-finally (yielding *nisa* from /*nisaj*/); before a consonant, the vowel-glide sequence undergoes coalescence to create a long mid vowel with the backness of the glide (yielding *niseet* 'I/you m. forgot' from /*nisaj+t*/). But this account leads us to expect verbs in which we find [*oo*] before the suffix, since there is no reason to exclude the possibility of verbs ending in /*w*/. The fact that all final weak verbs take [*ee*] before a suffix suggests that this vowel has been re-analyzed as a stem extender, rather than as the result of a general phonological process. In our terms, the function of this stem extender is to ensure that a distinction is maintained between the [+3] stems, which receive stress on their initial syllable, and the [−3] stems, which are stressed on the extender.

Also problematic for the phonological analysis are the geminate (or doubled) verbs:

(46) Geminate Verbs

	'send'	
3rd	dázz, dázzat, dázzaw	(+Ø, +at, +aw)
2nd	dazzéet, dazzéeti, dazzéetu	(+t, +ti, +tu)
1st	dazzéet, dazzéena	(+t, +na)

In these verbs, as in the final weak verbs, [*ee*] appears before the [−3] suffixes. We could ascribe the appearance of [*ee*] to the presence of a final glide, assuming underlying /*dazzaj*/ (parallel to *baddal* 'change'). However, we would then expect the third-person masculine singular (the unsuffixed form) to surface as **dazza*, by the same rule that deletes the final glide in *nisa* 'he forgot'. On the other hand, if we assume that the stem is either /*dazz*/ or /*dazaz*/ (with metathesis), we have no explanation for the appearance of [*ee*] before consonant-initial stems. While syllable structure constraints would indeed prevent the faithful realization of inputs like /*dazz+t*/, /*dazz+ti*/, we would expect these forms to be made pronounceable via more widespread processes of epenthesis or degemination:

(47) a. /dazz+t/ ⇒ dazzeet 'I/you m. sg. sent'
expected form: *dázzit via epenthesis (cf. /kitab+t/ ⇒ kitabit 'I/you m. sg. wrote')

b. /dazz+ti/ ⇒ dazzeeti 'you f. sent'
expected form: *dázti via degemination (cf. /baddal+at/ ⇒ badlat 'you f. sg. changed')

Thus, there is no obvious phonological account of the appearance of [*ee*] in geminate verbs. Another problematic case involves so-called hollow verbs, which historically had a glide as their middle radical:

(48) Hollow verbs

	'see'	
3rd	ʃáaf, ʃáafat, ʃáafaw	(+Ø, +at, +aw)
2nd	ʃífit, ʃífti, ʃíftu	(+t, +ti, +tu)
1st	ʃífit, ʃífna	(+t, +na)

In these verbs, the stem is monosyllabic, leaving no room for a stress difference between [+3] and [−3] stems. However, the two sets of stems are nevertheless distinct, with [+3] stems containing a long low vowel and the [−3] stems containing a short high vowel. Thus, assuming a single stem for all persons, we need to explain the realization of /*ʃáaf+ti*/ as *ʃífti* 'you f. saw', rather than **ʃáafti*. There is no clear phonological reason for shortening the stem vowel in this context since CVVC is tolerated in this language (Broselow, Chen, and Huffman 1997); cf. *xáabrat* 'she telephoned'.[14]

The array of facts above do not lend themselves to a single phonological analysis. However, we can describe them all as an effect of an imperative for morphological distinctness. In each verb type, we see a contrast between the base of suffixation in [+3] and [−3] forms. For bisyllabic stems, the [+3] base consists of either a stressed followed by an unstressed syllable, or a single stressed syllable, while [−3] bases have stress on their second syllable. Monosyllabic bases are of two types: final weak verbs and geminate verbs add a second syllable ([*ee*]) in [−3] forms, and this syllable bears stress; hollow verbs are monosyllabic in both [+3] and [−3] forms, but the single stressed vowel changes its quality in [−3] forms. We can assume a constraint enforcing non-identity between [+3] and [−3] bases (reminiscent of Alderete's (1999b) antifaithfulness constraints):

(49) [−3] Contrast:

A base bearing a non-third-person ([−3]) suffix must be distinct from the unmarked [+3] base in the identity of the stressed vowel.

This constraint compares the base of a [−3] suffix to the non-suffixed third-person masculine singular. Hollow verbs satisfy this constraint by changing the quality of the stressed vowel, while the other verb types satisfy it by means of locating stress on a different vowel. We can now view the exceptional stress

[14] Brame (1970) proposes an analysis for the counterpart verbs in Modern Standard Arabic whereby underlying ʔ*ʃajaf+ti*/ is transformed first to *ʃajf+ti*, with subsequent vowel-glide coalescence.

in *kitábIt* 'I/you m. wrote' as a result not of stress-epenthesis interactions, but of the desire to maximize contrast between the [+3] and [−3] forms:

(50) /kitab+at/ 'wrote, 3 f. sg.'	*Complex Coda	[−3]Contrast (base for comparison: kítab)	Stress Constraints	Syncope
a. kítabat				*!
☞ b. kítbat				
c. kitábat			*!	

(51) /kitab+t/ 'wrote, 1 sg./ 2 m. sg.'	*Complex Coda	[−3]Contrast (base for comparison: kítab)	Stress Constraints	Syncope
a. kítabIt		*!		
b. kítbIt		*!		
☞ c. kitábIt			*	

This is by no means a full account of Arabic stress and epenthesis (see Broselow (1992), Piggott (1995), Kiparsky (1999) for discussion of a broader range of data). But it does at least suggest an approach to the complex morphology of perfect stems.

4.6 Conclusion

An examination of the interaction of stress and epenthesis reveals a rich and complex array of facts, with epenthetic vowels sometimes patterning with underlying vowels, and sometimes disrupting the normal stress patterns, and both patterns sometimes coexisting within a single language. I have argued that Alderete's basic insight, that languages tend to avoid placing epenthetic material in prosodically prominent positions, allows us to account for many cases of apparently exceptional stress. Other cases may be accounted for by principles of maximization of morphological contrast. Based on the data here, it appears that a serial account of these facts is neither necessary nor desirable.

Appendix: Residual Issues

1 *Selayarese*

Though Selayarese does not have contrastive vowel length, vowels in open syllables show an increase in length under stress (Basri 1999). In (52), we see that while underlying vowels lengthen before a possessive suffix, epenthetic vowels do not; the epenthetic vowel in (52b) is followed by a geminate consonant (the phonetic realization of a glottal stop followed by a voiceless consonant):

(52) a. /sahala+ku/ ⇒ sahalá:ku 'my sea cucumber'

b. /sahal+ku/ ⇒ sahalÁkku 'my profit'

Stress falls on the epenthetic vowel in (52b), in violation of HEAD-DEP, because the alternative footing, in which the first two syllables constitute the main stress foot, would violate constraints against leaving a sequence of two syllables unparsed (see Broselow 2000 for a complete analysis).

Piggott (1995) argues that the failure of epenthetic vowels to lengthen under stress is evidence that these vowels are not present when the lengthening rule applies, supporting a serial analysis of epenthesis. However, an alternative analysis of these data has been proposed by Basri (1999). Basri argues that glottal stop insertion is preferred to vowel lengthening (NOLONGV ≫ DEPC) as a means of satisfying the requirement that stressed syllables be bimoraic. But glottal stop insertion is blocked in vowel-final stems by a higher-ranked alignment constraint requiring the right edge of the stem to coincide with the right edge of a syllable boundary (ALIGN-R(STEM,SYLLABLE)). In forms such as *sahalá:-ku* 'my sea cucumber', the alignment of stem-final [a] would be destroyed by insertion of glottal stop. But in *sahalÁ-kku* 'my profit', the right-edge alignment constraint will be violated no matter whether the bimoraic condition is satisfied by vowel lengthening or by glottal stop insertion, because the rightmost stem segment, [l], is not a possible coda. Therefore, the preferred option of glottal stop insertion is chosen. (Basri does not discuss why stressed vowels within a morpheme are lengthened; presumably, glottal stop insertion would be blocked by high-ranked CONTIGUITY.)

The addition of possessive suffixes provides the only environment in which epenthetic vowels can receive stress in native vocabulary, because these are the only consonant-initial suffixes that fall within the stress domain (see Basri, Broselow, Finer, and Selkirk 1997*a, b*, 1999). But the loanword data present a wider range of epenthesis sites. Epenthetic vowels within a stem (as in *karÁtu* 'card') are not followed by a glottal stop/geminate, but do in fact lengthen under stress, just like underlying vowels. This is consistent with Basri's account, but problematic for Piggott's.

In fact, it is arguable whether the gemination/glottal stop insertion seen before possessive suffixes is best analyzed as an effect of adding weight under stress, rather than a property peculiar to the possessive suffixes themselves (as Sirk (1988) shows, most South Sulawesi languages have two sets of possessive suffixes, -CV and -CCV, with the alternation frequently dependent on morphological rather than phonological factors). But in either case, the failure of vowels before possessive suffixes to lengthen does not provide a compelling argument for a serial analysis.

2 *Winnebago*

This section addresses some residual issues regarding Winnebago accent placement. Non-initial heavy syllables in Winnebago bear accent, so we find forms like *kiríina* 'returned', in contrast to forms like *hotaxí*. We can account for this by assuming that two additional constraints are active in Winnebago accent placement: a constraint requiring heavy syllables to bear accent (the accentual counterpart of WEIGHT TO STRESS), and a constraint forbidding accent on the initial syllable (arguably, the same constraint that accounts for the low pitch on Japanese initial syllables). The ranking illustrated below will derive the correct accentual patterns:

(53) /hootʃagra/ 'the Winnebago'	NOINITIAL ACCENT	OCP	HEAVYHEAD ACCENT	POST ACCENT	HEAD ACCENT
☞ a. {hoo}{tʃágra}			*		**
b. {hóo}{tʃágra}	*!	*			*
c. {hoo}{tʃagrá}			*	*!	*
(54) /kiriina/ 'returned'					
a. {kirii} ná			*!		*
b. {kiríi}ná		*!			
☞ c. {kiríi}na				*	

Remaining problems include binary/ternary alternations illustrated by the contrast below, in which stress falls in (55a) on the third and sixth syllables, but in (55b) on the third and fifth:

(55) a. hokiwároroké 'swing (v. intrans.)'

b. hakirúdʒikgádʒa 'after he pulls taut'

Following Hale (1985), I assume that these forms differ in their morphological structure, and that footing is sensitive to morphological constituency. Similarly, Hale argues that the following form demonstrates the necessity of incorporating reference to morphological structure in the analysis:

(56) hirat'át'aʃAnakʃÁna 'you are talking'

predicted form: *{hira}{t'át'a}{ʃÁna}{kʃÁna} (accent on 5th syllable)

Hale (1985): 2 metrical domains, {hira}{t'át'a} and {ʃÁna}{kʃÁna}

The problem of binary/ternary alternations illustrated above is of course independent of the question of whether stress/epenthesis interactions are best handled by serial or parallel accounts.

5

Reduplicative Economy[1]

WILLIAM IDSARDI AND ERIC RAIMY

5.1 Introduction

The unexpected interactions between reduplication and phonological processes were the direct motivation for the invention of Correspondence Theory (CT; McCarthy and Prince 1993, 1995*a*, etc.) as a component of Optimality Theory (OT; Prince and Smolensky 2004, McCarthy and Prince 1993, etc.). As uncovered in the pioneering study of reduplication, Wilbur (1973), phonological processes frequently do not apply normally when combined with reduplication. Rather, what is observed is surface over- or underapplication of the phonological processes. McCarthy and Prince analyze these departures from normal process application as instances of identity (faithfulness) between Base and Reduplicant. Extending such analyses to other areas of phonology, McCarthy and Prince (1995*a*) expand Correspondence Theory to cover input-output relations generally. As Correspondence Theory has evolved, it has gained a whole panoply of additional uses, such as: Input-Reduplicant relations (McCarthy and Prince 1995*a*), Output-Output relations within a paradigm (Benua 1995) and Base-Epenthesis relations (Kitto and de Lacy 1999).

The advent of Correspondence Theory introduces new constraint families (Max, Dep, Ident) but also significantly enriches the representations allowed within the theory. The original Containment model of Prince and Smolensky (2004) handled deletion through underparsing, and epenthesis through empty prosodic positions (symbolized there by an empty box), which were to be filled in at phonetic interpretation. For a variety of empirical and conceptual reasons, the Containment model was superseded by Correspondence Theory, which allows a much richer conception of deviation from

[1] This work was supported by a Fulbright fellowship awarded to William Idsardi while at the University of Toronto.

input-output identity. The difference in the two theories lies in the output of Gen, as in (1).

(1) Containment: Gen(input) → {outputs}
Correspondence: Gen(input) → {<input, output, relation>}

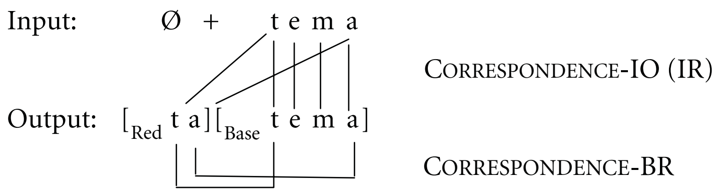
In the Containment model Gen produces a set of candidate output forms, each of which is an elaboration of the input form. Nothing of substance is removed, rather items which will ultimately remain unpronounced are included in the "output" so that they may still condition phonological processes. In contrast, the Correspondence model allows for arbitrarily different output forms, but includes a function within each "output form" indicating the correspondence relation between input and output.

In both theories there is an equivocation on what "output" means. Hale (2000) identifies the core of this issue as whether the "output" of an OT grammar is a purely phonetic representation that has already been transduced (transformed from discrete abstract phonological features to less abstract more continuous phonetic features) or is a phonological representation that has to be transduced at a later point in the derivation. While this issue is not crucial to the issues discussed in this chapter, it is one that must be addressed before a definitive evaluation of the success and merits of OT can be accomplished. For the purposes of this article, we will assume that the "output" of an OT grammar is a phonological one that must be further converted to a phonetic representation later in the derivation. Hale (2000) provides interesting discussion of the ramifications of this position.

The purpose of this chapter is to examine some of the consequences of Correspondence Theory. We believe that Correspondence Theory is far too powerful, and as a result it is analytically uninsightful and computationally implausible (see Idsardi 2006 for discussion). We offer as an alternative the representations for reduplication proposed in Raimy (2000*a*). These representations have a greater affinity with the original Containment model of Prince and Smolensky. In Raimy's model, an abstract phonological representation is calculated which is then submitted to the phonetic implementation component.

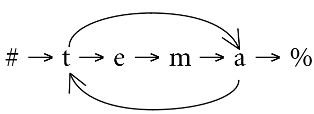
Specifically, Raimy (2000*a*) clarifies how the information about temporal precedence is represented phonologically. Temporal relations are best encoded in phonological representations with a particular data structure, namely a linked list. The simplest temporal patterns can be represented with "linear" linked lists, those in which each segment leads to exactly one other segment. Once we identify this data structure as appropriate for simple cases, we can then ask a more sophisticated question: What could we represent by relaxing the linearity condition? That is, what if some non-linear temporal precedence relations were allowed? What if some segments were specified as having two or more possible continuations? Could we capture over- and underapplication more insightfully with these slightly enriched representations?

To illustrate the difference between these approaches, let us consider how a hypothetical[2] reduplicated form *ta-tema* would be represented in the two approaches. In CT, there will be an input, an output, and two correspondence relations, Correspondence-IO and Correspondence-BR, as in (2).

(2)


The [ta] portion in the output has a separate representation from the [tema] portion, but the two are related by Correspondence relations. First, they have correspondence relations to each other, Correspondence-BR. Second they share correspondence with segments in the input, Correspondence-IO.

In contrast, we show Raimy's representation in (3).

(3)


In (3) there is no separation of [ta] from [tema], they use the same pieces of phonological structure. The portion [ta] simply repeats certain portions of [tema]. There are two continuations from the /t/, one to /a/ and one to /e/, and there are two continuations from /a/, one back to /t/ and one to the end of the word. Once this structure is linearized, we will obtain the surface form [tatema].

In the rest of this chapter we will briefly show how the Raimy model of reduplication works. In particular we will show how this system analyzes over- and underapplication facts and relates these to other phenomena in an insightful fashion. But, most importantly, we will show that Raimy's system is representationally economical. By this we mean that there is a natural evaluation metric for the complexity of representations in Raimy's system, and the right results come out when the simplest representation for reduplication

[2] This is basically the pattern of reduplication in Semai (Diffloth 1976).

through "looping" is chosen. In contrast, CT is too powerful because it allows arbitrary deviation between Base and Reduplicant, and it allows arbitrarily complicated correspondence relations among several representations. Simply put, CT allows too large a space of possible analyses for reduplication. Since the child must settle on the right analysis, CT requires too much data inspection on the part of the child. Consequently CT analyses are much less learnable than Raimy-style analyses.

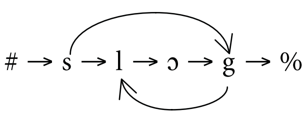
5.2 Representing reduplication in Temiar

To illustrate the differences between the two approaches, we will consider some facts about reduplication in Temiar (Benjamin 1976 and most recently Gafos 1998, see references in Gafos for previous analyses). The continuative is formed in Temiar by repeating a portion of the word, as in (4).

(4) /Continuative + slɔg/ >>> [sglɔg]

If our job is to relate /slɔg/ to [sglɔg] as perspicuously and parsimoniously as possible with the resources in the Raimy model, we are forced to the analysis in (5).

(5)



We start at the /s/, jump to the /g/, come back to the /l/ and finish off the word as normal. The continuative morphology does not add any new segments but does condition the addition of two new temporal relations, the jump from /s/ to /g/ and the loop back from /g/ to /l/. Raimy (2000*a*) explains that the added temporal relations take precedence over those that are already present (this is a principle of Universal Grammar). The structures created in this way are also interpreted as economically as possible, so that the smallest output consistent with the maximum number of consistent temporal relations is chosen as the output, in this case [sglɔg]. Notice that the link [s → l] present in the underlying form is not used during linearization. Omissions of precedence information like this arise when two paths emanate from a segment, but only one path leads into the segment. In that case one of the paths must be sacrificed in pronunciation, and it is always the underlying path that is sacrificed in order to use the newly created path. In general the interpretation of the phonological structures optimizes for the following considerations: (1) the output must be asymmetric (this is inviolable), (2) no new precedence links are added (also inviolable), (3) morphologically added information is used first, and (4) the shortest possible output is generated (Raimy 2000*a*).

Raimy demonstrates that the Continuative morphology of Temiar is accomplished by the processes informally stated in (6).

(6) a. Add a link from the *last* segment to the *first onset* segment.

b. Add a link from the *first* segment to the *last* segment.

The representation that results from the addition of the precedence links indicated in (6) to a base compactly and explicitly stores the information necessary to pronounce both forms. The continuative is built out of the basic form by adding new information to it, and segments are reused whenever possible.

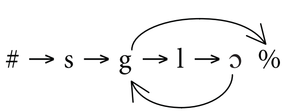
This new understanding of the phonological representation of temporal information has two major advantages: (1) it is not tied to particular surface properties or constructions and (2) it has an obvious and natural evaluation of markedness in the size of the structures required for representations and process statements. That is, the representation is both abstract and economical. It is these two properties that make this an excellent structure from which we can learn more about the way in which phonological relations are represented and manipulated mentally.

5.3 Economy of representation

Raimy's analysis of reduplication offers the beginnings of an explanation of how reduplication can be recognized and processed by the child, and why reduplication is a word-formation strategy favored by many languages. The child can recognize repeated chunks in short-term memory and build up an expectation of how the form will continue. When the child hears [sglɔg], there are two possible analyses available to him; shown in (7).

(7) [sglɔg] is heard, child constructs:

a. # → s → g → l → ɔ → g → %

b. 

The reduplicated representation in (7b) is more economical than the non-reduplicated one in (7a) because it has one fewer segments and the same amount of precedence links. If this hypothesized representation holds up under further data it will be maintained by the child. This suggests that words with common subsequences may be represented as reduplicative even in languages without reduplication (such as English 'banana', (8)).

(8) # → b → a → n → a → %

Whether the representation in (8) is the one a child adopts while learning English will be affected by other considerations such as how the representation relates to metrical information (for example, the surface difference of stress between the last two syllables may prevent the child from adopting this representation). This seems to be attested in Manam, see Buckley (1998) for the relevant facts and Fitzpatrick and Nevins (2004) for an analysis along the lines suggested here.

5.4 Economy of computation

Both the Raimy model of reduplication and the Correspondence model of reduplication require computations to be performed on "input" forms to produce correct "output" forms. The Raimy model presents a case where there is only a small difference between the input and output representations. This difference between input and output is minimized because linearization only produces output forms that consist of links present in the input form. No new precedence relations are ever added during linearization, and only rarely are precedence links lost in the output (as in the above Temiar case). This economy of computation is what makes the Raimy model similar to the Containment model of Prince and Smolensky. Both of these models limit the amount of potential computation because the possible output representations that are produced are directly constrained by what information the input representations already contain. Correspondence models do not constrain the possible outputs that can be produced by a process. Outputs that freely add or delete any kind of structure are considered in Correspondence models and this increases the cost of computation dramatically.

Another aspect of computational economy present in the Raimy model is the statements that are required to add the links that produce reduplicative loops. These statements simply consist of specifications of what two segments stand in the relationship of precedence. Consider again the informal statement of Temiar reduplication in (6), presented below in the formal representation proposed in Raimy (2000*a*) (9).

(9) a. Add a link from the *last* segment to the *first onset* segment

begin → *end*   *begin*:   _ → %

end: # . . . X
            |
          onset

b. Add a link from the *first* segment to the *last* segment

begin → *end*   *begin*:   # → _

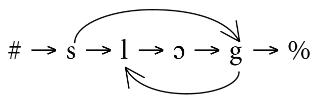
end: _ → %

The formalism in Raimy (2000*a*) specifies that each precedence link has a *begin* and *end* specification. These specifications indicate how the added precedence link will concatenate with the base. All forms have a beginning and an ending and, presumably, these positions are the two easiest locations to identify. All positions of segments in a formative are in ultimate reference to the beginning or ending of a form, thus all *begin* and *end* specifications of a precedence link will have the notion either "first" or "last" inherently present in them.

Total reduplication results from a precedence link that states "the last segment precedes the first segment." This can be considered the least complicated specification of a precedence link since it only requires reference to the first and last segment of a form. Support for this conclusion is the fact that total reduplication is the most prevalent type of reduplication (Moravcsik 1978). The coincidence between the formal simplicity of the statement of total reduplication and its prevalence in the world's languages creates the basis of a natural markedness metric. As more information is added to the link statements, the resulting reduplication pattern should be less prevalent. This appears to be true since less frequent but common reduplication patterns, such as prefixing CV/CVC, require the additional specification of some sort of prosodic position (nucleus, onset, C, etc.) in the statement of one part of the precedence link. Infixing reduplication which appears to be still rarer requires additional prosodic information in both parts of the statement. The claimed rarity of suffixing reduplication in comparison to prefixing reduplication can be explained by the fact that prefixing reduplication will always include some reference to "first" (more generally the beginning of the form) while suffixing reduplication will always include some reference to "last" (more generally the end of the form). If the scan through a phonological representation starts at the beginning of the form, precedence links that can be concatenated without having to scan through the entire form will be favored over links that must reach the end of the form before they can be discharged—another computational economy. Patterns of reduplication like Temiar require two distinct links to be added and thus are more marked than other patterns that require only the addition of a single link.

We can understand how over- and underapplication effects are produced by considering how segments with multiple temporal relations are viewed by processes. Consider a hypothetical case: Temiar with word-final devoicing (WFD). How does the structure in (5), repeated in (10), react to the application of WFD?

(10)



There are three possibilities to consider. First, WFD could be delayed until after linearization, for example if WFD were a postlexical rule. In this case WFD applies normally: [sglɔk]. Second, the fact that /g/ is at the end of the word according to one of its precedence links could be sufficient to allow WFD to apply (a "contamination" view of the word-final environment); this produces overapplication as /g/ → [k], [sklɔk]. Finally, the fact that /g/ is not consistently word-final (the two precedence links differ on this point) could block the application of WFD, yielding underapplication [sglɔg]. It is the representational discovery of the ambiguous status of /g/ that allows for a unified explanation of both under- and overapplication. See Raimy (2000*a,b*) for analyses of various actual over- and underapplication effects and discussion of how backcopying effects are accounted for within this system.

5.5 Extensions

A natural consequence of the addition of a new device to a theory is the possible utility of it in other situations. There are two immediately possible extensions[3] of the Raimy model that do not relate to reduplication.

The first possible extension of Raimy (2000*a*) is to reanalyze deletion in phonological representations not as the actual removal of a segment but instead as the addition of a jump link. This approach to deletion can mimic underparsing analyses from Prince and Smolensky. (11) presents a possible approach to deletion in Tiberian Hebrew where instead of deleting the vowel in question, a link that skips over it is added (11b), which then allows

[3] Both of the following extensions are utilized in the analysis of reduplication in Tohono O'odham and Indonesian in Raimy (2000*a*).

spirantization (11c) to apply without ordering restrictions. In essence the extra representational power can stand in for derivational opacity in this instance.

(11) a. # → k → a → t → a → b → u → %

b. Deletion (Jump)

→ k → a → t → a → b → u → %

c. Spirantization

→ k → a → θ → a → v → u → %

d. Linearization

→ k → a → *θ* → v → u → %

This potential analysis preserves the opacity effects in that the vowel is present in the representation long enough to trigger spirantization and it is removed only when the form is linearized. This approach to deletion can also handle base-truncation effects (Benua 1995) and provides an explicit account of what truncation morphology is. The connection between reduplication and truncation as argued for by Benua is preserved in this model because over- and underapplication effects are both derived as cases of phonological opacity involving segments with multiple links.

Another possible area of extension is the representation of geminates. A natural representation of geminates would be to have a segment "loop back" onto itself as in (12).

(12) # → t → a → k → i → % Agta takki 'leg'

This approach to the representation would immediately connect geminate blockage effects (Schein and Steriade 1986) with over- and underapplication effects. Both of these phenomena would be united under the understanding that it is the environmental ambiguity that a single segment appears in that is the source of their unusual behavior. It must be noted that the present representation of gemination as the multilinking of a single melody to multiple x slots is not ruled out in the Raimy model and only future research will indicate whether both types of geminate representations are required or if one type can explain all geminate behavior.

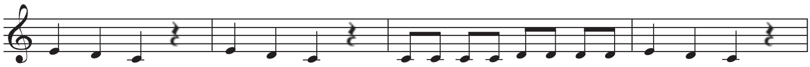



FIGURE 5.1. Hot Cross Buns[4]

5.6 Repetition in musical cognition

Precedence is a basic notion in phonology and if the Raimy model of precedence is correct we should hope to find converging evidence from other similar areas of cognition. A discussion of the learning of a musical tune by Bamberger (1991) provides strong supporting evidence for the representation of repeated material as proposed in the Raimy model.

Bamberger (1991) discusses an experiment where a child is asked to build the melody for "Hot Cross Buns" using what are called Montessori Bells. The melody for "Hot Cross Buns" is presented in Figure 5.1.

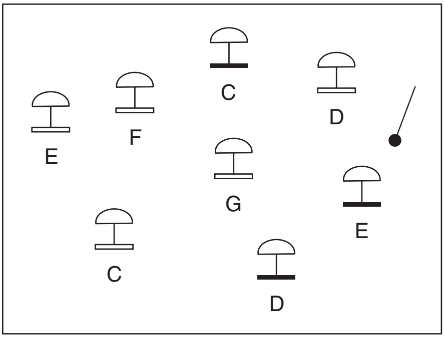
Montessori Bells are an important aspect of this experiment because of their ingenious design. Most musical instruments give some indication of what pitch they can produce through physical attributes. For example, we expect a smaller shorter instrument such as a piccolo to have a higher pitch than the larger and longer flute. A correspondence between pitch and size is something that we learn very quickly and it is a connection that the Montessori Bells deliberately avoid. Montessori Bells are built specifically to be identical in appearance but to still have different pitches. Thus, a C bell and a D bell can only be distinguished by striking and listening to them and cannot be distinguished by looking at them.

The particular Montessori Bells used in this experiment covered the C-major scale starting at middle C (eight bells) and the entire chromatic scale starting at middle C (thirteen bells).[5] This provides the child enough resources to build the tune for "Hot Cross Buns" which is found in Figure 5.1. The "search space" for the child in the experiment discussed by Bamberger is presented in (13). (13) is considered the search space because all of the notes needed to construct "Hot Cross Buns" are present in (13).

[4] It is unclear for percussive instruments such as the Montessori Bells whether the duration of the notes (quarter notes for this example) is relevant because it is the timing of the striking of the bells that is important and not how long the bells actually ring.

[5] There is another dimension of color to the bells provided to the children in this experiment. The chromatic bells had brown bases while the C-major bells had white bases. Bamberger discusses the special use of color in the learning task but this aspect is irrelevant to the point at hand and will not be discussed.

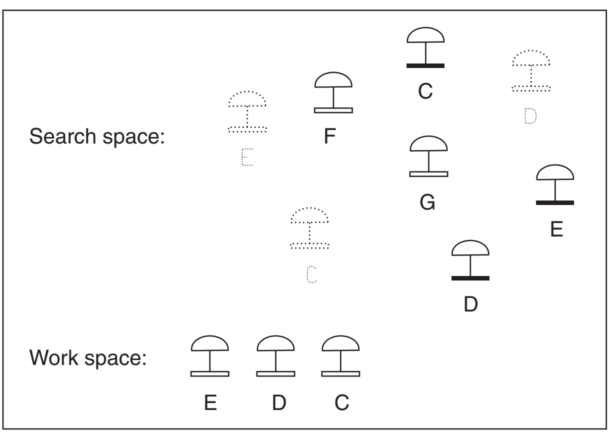
(13)



Bamberger (1991: 132, figure 7-2)

The interesting and relevant part of this experiment is how the child built the tune in Figure 5.1 out of the bells given in (13). The subject was given an E bell to begin and was allowed to proceed in building the tune from there (the child already knew the tune from a previous experiment). The child built the first measure by randomly choosing bells and striking them, listening to see if it matched the tune. Once a matching bell was found it was placed next to the previous note. The first measure of the tune was completed and the child had three bells sitting in front of him separated from the search space as seen in (14).

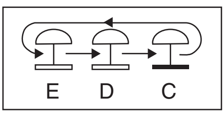
(14)



Bamberger (1991: 134, figure 7-5)

Now the interesting thing is what happens next. Instead of searching for another E bell to continue the tune, the child immediately used the three bells present in the "work space" to produce the second measure. The child recognized that the first and second measures of HCB are identical and recycled the present resources to reflect this fact. Most striking about this behavior is that a previous experiment where the child used a computer program to build HCB did not allow the child this possibility. There was no way to tell the computer to repeat a sequence. It appears that given the opportunity to use the notion of "repeat" in a representation, the child does. Bamberger (1991) provides the representation in (15) for the present state of HCB in the experiment. The similarity between Bamberger's informal representation in (15a) and a representation of this behavior in the representations proposed by Raimy (2000*a*) is striking.

(15) a.



b. # → E → D → C → %

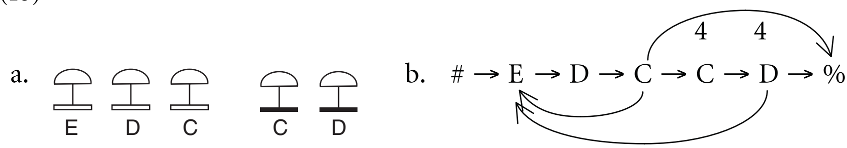
Bamberger (1991: 135, figure 7-6)

The behavior of the child after these first two measures are constructed indicates that more than surface similarity between two entities is required for them to be equated as the same in a representation. Consider (15): this representation contains all of the bells that are required to play HCB. From (15) we only need to add the four repetitions of C, the four repetitions of D, and then return to the beginning. All of these notes are contained in (15). Interestingly, the child does not recognize that the final C note in measure two of HCB is the same note as the one that begins measure three. The child discussed by Bamberger actually explicitly says that the final C note in measure two is not the same note that begins measure three when asked about it.

At this point in the child's construction of HCB, the child is searching for the first note of measure three by playing through the first two measures and then striking a single bell four times to see if it "fits" the melody. This indicates that rhythmic information plays a role in the determination of "identity" in a representation and explains why the child does not recognize the single note of C in the first and second measure as the same note as the repeated C note in measure three.[6] Eventually, the child finds another C bell and another D bell and creates the sequence of Montessori bells in (16a).

[6] That more than simple surface resemblance is considered when building representations will allow languages to build reduplicative structures into representations in an "intelligent" way that allows Manam to build reduplication into forms that repeat the final syllable (i.e. ragogo is ragoRED) as claimed by Buckley (1998) and Fitzpatrick and Nevins (2004) but to not build a reduplicative structure in a form like *banana* in English. Whether repetition is adopted in a representation is affected by the language-specific information based on criteria such as segmental rules, metrical considerations, morphology, etc.

(16)



Bamberger (1991: 138, figure 7-9)

The organization of bells in (16a) is sufficient to produce HCB. The child first runs through the EDC group two times and then plays four repetitions of C and D. To complete the tune, the child now immediately goes back to the first set of bells and runs through the EDC sequence with no hesitation. This is important because it indicates that while rhythm may prevent two identical notes from being recognized as identical, temporal distance may not interfere with this type of judgement in a representation. The final representation of the tune constructed by the child within using the representations in Raimy (2000*a*) is in (16b). Note that the numbers above the second C and D note are shorthand for four loops back to the note itself.[7] This representation is completely in line with the Raimy model of reduplication, and consequently the discussion by Bamberger (1991) of children learning melodies with Montessori Bells supports the representational proposals of Raimy (2000*a*). This further supports Lerdahl and Jackendoff's (1983) parallels between music and language.

5.7 Representational profligacy in Correspondence Theory

As discussed in the introduction, Correspondence Theory analyzes reduplication by establishing a Base (B) and a Reduplicant (R) in the output string, along with separate Correspondence relations between B and R and input and output. The whole information structure—input, output, where B is, where R is, Correspondence-IO, Correspondence-BR—constitutes the phonological structure produced by Gen and submitted to Eval. For convenience, let us call this the GENOUTPUT. There is no requirement that a given surface form has a unique GENOUTPUT and in fact it may be the case that there is never a unique GENOUTPUT for any reduplicative structure. To illustrate this point, we will consider GENOUTPUT structures for the previously discussed Temiar form in (4) that have either been proposed in the literature or must be considered by Gen based on other existing OT analyses of reduplication.

[7] Some indication of "rest" should probably be added to (16b).

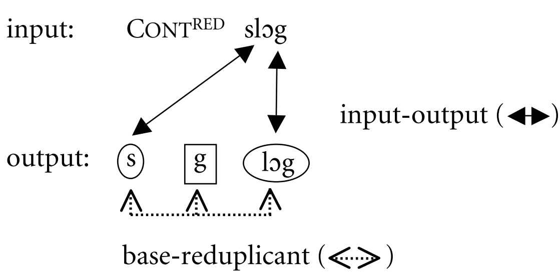
We can begin by considering the analysis of Temiar offered by Gafos (1998). Although Gafos does not explicitly discuss what the base is or how it is calculated, we can determine what his assumptions are by looking at the tableau in (17) (Gafos 1998: 520).

(17)

	CONT^RED, slɔg	*PREFINAL-V	MARKEDNESS	MAX-BR
	a. sɔ.lɔg	*!	*****	***
	b. s.**1g**.lɔg		******!	**
☞	c. s.**g**.lɔg		*****	***

The important aspect of the tableau in (17) is how the violations of MAX-BR are calculated. Since the winning candidate (17c) only reduplicates a single consonant (/g/ in bold) and there are three violations of MAX-BR we can conclude that Gafos considers the stem in the input to be coextensive with the base in the output. This produces the GENOUTPUT structure in (18) as the winning candidate.

(18)



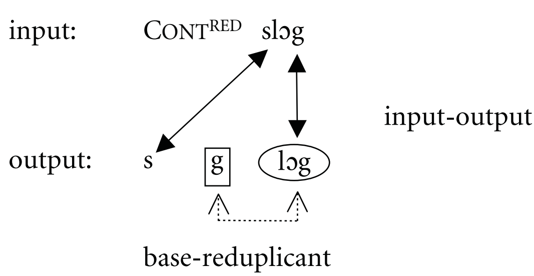
The GENOUTPUT structure in (18) presents the relevant correspondence relationships that are used to calculate input-output Faithfulness and base-reduplicant Faithfulness. The input level indicates that a stem /slɔg/ and an abstract reduplicative morpheme associated with Continuative Aspect are present. The output level contains the phonological string /sglɔg/. The base structure is indicated by enclosing all segments considered to be the base in ovals. There are two regions of base in (18), /s/ at the beginning of the output and /lɔg/ which is at the end. Splitting the base is the /g/ which is the reduplicant. The reduplicant in (18) is marked by being enclosed in a rectangle. The correspondence relations of input-output and base-reduplicant are indicated by the arrowed lines. The solid lines with solid arrow heads indicate input-output correspondence and the dotted lines with simple arrow heads indicate base-reduplicant correspondence. Input-reduplicant faithfulness is suppressed in this example because it is not relevant to Gafos's analysis.

The GenOutput structure in (18) is not the only possible one for the particular Temiar example. In fact, the analysis proposed by Gafos (1998) violates claims made by Urbanczyk (1996, 2001, 2006) about what possible relationships between base and reduplicant can occur in Gen. Urbanczyk (1996: 272, 2001) proposes the *adjacent string hypothesis* which states that the "B[ase] is the string adjacent to R[eduplicant] such that it begins at the tropic edge." The *tropic edge* is defined as "...immediately follow[ing] R if R = prefix or immediately preced[ing] R if R = suffix" (1996: 272, 2001). Following from these points, Urbanczyk (1996: 273, 2001) states that:

> An implicit feature of A[djacent] S[tring] H[ypothesis] is that there is a direct relationship between the placement of RED and the segments which constitute the base. So, as Gen may freely place the reduplicant anywhere in the word, the tropic edge, and hence the base will be determined accordingly.

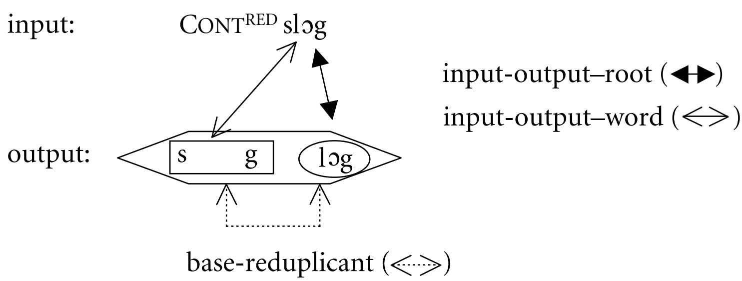
Urbanczyk's restrictions on how the base is calculated by Gen rules out the GenOutput structure in (18) because the base cannot be on both sides of the reduplicant. Accordingly, a different GenOutput, based on (18), that follows Urbanczyk's restrictions on the base is presented in (19).

(19)



The difference in GenOutput structures between (18) and (19) is that in (19) the /s/ in the output is not part of the base. We must also assume that reduplication in Temiar must be prefixation in order for the base to be to the right of the reduplicated /g/. This ensures that there is a Faithfulness relationship between the reduplicant and base which can determine the correct segmental content of the reduplicant.

A yet different GenOutput structure is consistent with Urbanczyk's restrictions on the calculation of the base if proposals in Struijke (2000, 2002) are considered. Struijke (2000, 2002) modifies McCarthy and Prince's (1995*a*) extended model of reduplication by replacing the notion of input-reduplicant faithfulness with *word faithfulness*. Word faithfulness calculates input-output Faithfulness on the entire output string of segments regardless of base-reduplicant association. Given this modification, the /s/ in the GenOutput structure in (19) can be associated with the reduplicant producing a new GenOutput structure in (20). The additional input-output–word Faithfulness calculation is included in (20) by the additional hexagon which indicates the domain of Word Faithfulness. The input-output–word correspondence is indicated by the plain double-arrowed line.

(20)


The novel aspect of the GenOutput structure in (20) is that although /s/ is part of the reduplicant, it is faithful to the input structure (and not the base in the output) through word faithfulness. All input-output-based Max constraints are thus satisfied since all segmental content in the input appears in the output as the model of Struijke (2000, 2002) outlines. Two possible advantages of this approach are that (1) the output is completely parsed into either a base region or a reduplicant region and (2) the reduplicant region for a triconsonantal form like /slɔg/ is now coextensive with the reduplicated region in a biconsonantal form /kɔ̄w/ >> /kwkɔ̄w/.

The theories of the base considered so far all assume that there is a universal static generalization as to what the base is for a reduplicant in a GenOutput structure. However, Kitto and de Lacy (1999) propose a model of epenthesis which claims that there is a correspondence relation between an epenthetic segment and a segment present in the input. The segment that corresponds with the epenthetic segment is thus the base in this correspondence relationship. Most relevant to the present discussion is the claim that the base for epenthetic consonants is not determined by a static function but instead is determined *dynamically* via constraint interaction. Directional effects in epenthesis patterns where an epenthetic segment derives some or all of its features from a segment in the input are produced through constraints that indicate whether the base for epenthesis may occur to the left or the right of the epenthetic segment in the output. Since direction of epenthetic copying is now determined by constraint interaction, patterns where epenthesis copies from the right in some circumstances but the left in others can be described. Consider the Faroese example discussed in Kitto and de Lacy (1999) presented below in (21). (21a) presents the generalization of the pattern and (21b, c) show tableaux that indicate how the appropriate constraints interact to encode the generalization in (21a).

(21) a. (1) Copy from the left if [i] or [u] precedes:
e.g. [si:j̱ur] 'custom', [hyuw̱ir] 'skins'
otherwise
(2) Copy from the right if [i] or [u] follows:
e.g. [so:j̱in] 'boiled', [mæaw̱ur] 'man'
otherwise
(3) Do not epenthesize

b.

	/o_i/	BE-IDENT-F	COPY-LEFT	COPY-RIGHT
	$o_1\underline{w}_1i$	*!		*
☞	$o\underline{j}_1i_1$		*	

c.

	/i_u/	BE-IDENT-F	COPY-LEFT	COPY-RIGHT
☞	$i_1\underline{j}_1u$			*
	$i\underline{w}_1u_1$		*!	

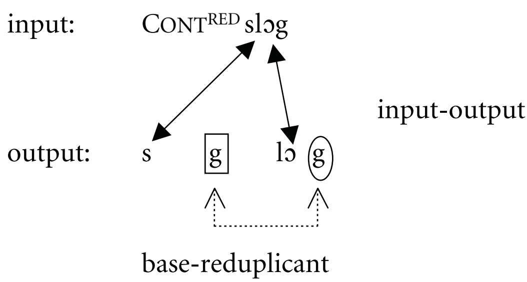
The BE-IDENT-F constraint in (21b) enforces featural similarity (specifically [high] for this case) between the epenthetic segment and its base. This is seen in the first candidate in (21b) where the epenthetic segment /w/ has been inserted but there is no feature [high] in the base (/o/ in this case). Since the candidate that copies from the right violates the IDENT constraint, the secondary pattern of copying from the left may emerge. The tableau in (21c) shows that copying from the right takes precedence when the IDENT constraint is satisfied by both candidates.

While Kitto and de Lacy (1999) only explicitly discuss base-epenthetic relations, they are clear on the wider ramifications of their proposal which affect possible GENOUTPUT structures. In their conclusion they state:

> … the Base cannot be identified by a "static" mechanism, but is instead "dynamic"—the location of the Base can change in different environments. In effect, the identification of the Base reduces to constraint interaction. This opens up the possibility that the Base of reduplication is similarly determined. (Kitto and de Lacy 1999: 22)

Following this proposal leads to yet another possible GENOUTPUT structure that must be considered for the Temiar reduplication pattern. This structure is presented in (22).

(22)



(22) presents a GENOUTPUT structure where the base for the reduplicant is the single segment /g/. This base-reduplicant parsing minimizes the violation of MAX-BR so this parsing is presumably optimal under some high ranking of that constraint. Determining which base-reduplicant mapping is desired for particular reduplication patterns in specific languages requires the addition of input-base constraints that evaluate particular base-input mappings. Presumably this set of constraints would interact to determine whether the base should be to the left or the right of the reduplicant and what size the base should be (minimize base, maximize the similarity between base and input, base should be a minimal prosodic word, etc.). This dynamic approach to determining what the base is in reduplication increases the number of candidates produced by Gen that must be evaluated. The unlikely but possible reduplicant-base mappings where /g/ is the reduplicant (as above in 22) but the base is /s/ or /l/ or /ɔg/, etc. or any other number of regions of the base must now be added to the candidates produced by Gen.

Excursus: The value of dynamic bases

While the dynamic theory of the base for reduplication appears to only complicate the analysis of reduplication patterns, the strict static interpretation of the base in reduplication as espoused by Urbanczyk (2001, 2006) runs into empirical problems when certain cases of reduplication are considered. (23) presents a case of reduplication from Indonesian discussed by McCarthy and Prince (1995*a*) with data from Uhrbach (1987).

(23)	pukul	pukul-məm-ukul	'to hit (reciprocal)'
	tari	tari-mən-ari	'to dance (reciprocal)'
	hormat	hormat-məŋ-hormat-i	'to respect (reciprocal)'

The importance and relevance of this pattern of 'interposing' reduplication is that the base and reduplicant are separated by the prefix /məŋ/. The separation of the base and reduplicant violates the *adjacent string hypothesis* as proposed by Urbanczyk (2001, 2006). Importantly, reduplication in Indonesian appears to be a case of total reduplication of the root[8] and this fact indicates that Gen must be able to somehow restrict copying to the input root in this pattern. This effect follows naturally if we assume that the base is calculated in the output via constraint interaction. The relevant question is what constraints determine the base. By considering other constraints already present in the literature on OT, we find extremely likely and useful candidates for relevant constraints. For the Indonesian example in (23), the relevant constraint is Base = Root[9] which will exclude affixes from participating in base-reduplicant correspondence if ranked high enough. Consider the tableau in (24) which provides an analysis of the interposed reduplication pattern in Indonesian.

(24)

	RED + məŋ + hormat + i	Faith-BR	Base = Root	Base = Stem	Alignment	ASH
☞	a. **hormat**-məŋ -hormat-i			****		*
	b. **hormat**-məŋ -hormat-i	*!**	***	*		
	c. məŋ -**hormat**-hormat-i			****	*!	
	d. **məŋ-hormat**-məŋ-hormat-i		*!**	*	*	

[8] We will ignore the interaction of nasal substitution and reduplication which complicates the first two examples in (23). See Raimy (2000*a*) for the analysis of the nasal substitution facts and the general interposing pattern of reduplication in Indonesian.

[9] The form of this constraint is based on Kager's (1999: 220) use of RED = STEM as part of the analysis of Diyari.

The function of each cover constraint in (24) is fairly intuitive. Faith-BR covers both Max-BR and Dep-BR and other possible deviance from perfect copying. Base = Root is one of the "base-specific" constraints that interacts to determine what the base should be in an output string and indicates that the base in the output should consist solely of the root. For this pattern of reduplication, Base = Root is undominated with respect to other "base-specific" constraints. Base = Stem is another "base-specific" constraint that indicates that the base should be the root plus any and all affixes associated with the root. Crucially, Base = Root dominates Base = Stem. Alignment is a cover constraint for the multiple constraints that determine the linear order between the root, affixes, and RED. Since /məŋ/ is a prefix, part of Alignment is a constraint that requires this affix to be to the left of the root. Since RED in this pattern is also a prefix, we must assume that the Alignment constraint that requires RED to be a prefix is ranked above /məŋ/ being a prefix which would give the RED + Affix + Root ordering in the output. Finally, ASH is a constraint that requires RED and base in the output to be in contact (this is similar to Raimy and Idsardi's (1997) *Gap constraint). Urbanczyk's definition of ASH can be adopted if it is treated as a violable constraint with violations occurring when either base and reduplicant are not contiguous or if the direction of copying violates Marantz's generalization (Marantz 1982).

Candidate (24a) is the most harmonic given the particular constraint ranking since there is complete identity between the base (underlined) and the reduplicant (in bold), the base coincides with the root, and the specified linear ordering of the affixes in the output is respected. Candidate (24b) is not as harmonic since the base calculated in the output contains more material than just the root. The affix /məŋ/ has been parsed as part of the base and this creates violations of Faith-BR and Base = Root which causes this candidate to be less harmonic. The lesser violation of the lower-ranked constraints, Root = Stem and ASH, does not save this candidate. Candidate (24c) correctly parses the base as only the root but places /məŋ/ outside of the reduplicant in the output which violates the relevant Alignment constraints causing it to be less harmonic. (24d) shows that adding material to the base in the output beyond the root and consequent additional material in the reduplicant does not produce a more harmonic candidate. The high-ranked Base = Root eliminates any candidates that do have the base and stem coextensive.

From this discussion, we can see that adopting Kitto and de Lacy's proposal that the base is determined dynamically in reduplication provides some immediate benefits. The contradiction between the analysis of Indonesian offered by

McCarthy and Prince (1995*a*) is reconciled with Urbanczyk's (1996) proposed restrictions on the calculation of the base if we adopt the standard mode of explanation within OT that generalizations result from the interaction of violable constraints. By transforming the *adjacent string hypothesis* (ASH) into a violable constraint, the Indonesian interposing reduplication pattern is no longer a counterexample to ASH but is instead a case where some higher-ranked constraint causes a violation of ASH–the standard state of affairs within OT.

The dynamic view of the base also allows prosodic circumscription effects in reduplication patterns to be directly accounted for. Consider the pattern of reduplication in Samoan (Kenstowicz 1994*b*).

(25)	táa	ta-taa	'strike'
	nófo	no-nofo	'sit'
	alófa	a-lo-lofa	'love'
	saváli	sa-va-vali	'walk'

The generalization offered by Kenstowicz (1994*b*: 635) of this reduplication pattern is that a light syllable is prefixed to the trochaic foot independently required for penultimate stress assignment. This reduplication pattern is succinctly captured in the dynamic base theory by simply specifying the base as "the main stressed syllable." Complementing this generalization with a high ranking of the ASH constraint creates the tableau in (26) which formalizes this analysis.

(26)

	RED + savali	Base = σ′	Faith-BR	ASH	Stress
☞	a. sa-**va**-váli				
	b. **sa**-sá-vali				*!
	c. **va**-saváli			*!	
	d. **sa**-saváli		*!*		
	e. **sa**-saváli	*!			

The tableau in (26) shows how the interaction of the constraints that capture the stress pattern of Samoan (the cover constraint Stress), the constraint that specifies what the base is (Base = σ′ "base equals syllable with main stress"), Faith-BR and ASH all cooperate to produce the correct surface pattern of reduplication. Candidate (26a) satisfies all of these constraints in that there is penultimate stress, the base is this stressed syllable, base and reduplicant are directly next to each other satisfying ASH and the entire stressed syllable is copied. All other candidates in (26) deviate from satisfaction of one of these constraints either misplacing the stress (26b), violating ASH (26c), copying the wrong string of segments (26d), or finally choosing the wrong base (26e). This analysis has the advantage of utilizing the surface-apparent stress pattern as the main determinant of what the reduplicant should be through the specification of what the base is. No type of prosodic circumscription or "misalignment" of the reduplicant is necessary. All the learner has to do is identify what the base is. This task is aided in Samoan by the fact that the base is the syllable with main stress, presumably a highly salient target.

A final advantage of the dynamic theory of the base is that it provides a principled explanation for the lack of Hamilton-Kager Conundrum (McCarthy and Prince 1999) effects in reduplication. The Hamilton-Kager Conundrum is the moniker for the observation that although segmental backcopying effects are found in reduplication, there are no attested cases of the backcopying of a reduplicative template resulting in the truncation of a base. The tableau in (27) presents this hypothetical situation.

(27)

	/RED + tilparku/	RED = MinWd	MAX-BR	MAX_{IO}
☞	a. tilpa-tilpa			***
	b. tilpa-tilparku		*!**	
	c. tilparku-tilparku	*!		

As McCarthy and Prince (1999) discuss, by ranking a templatic constraint and MAX-BR above MAXIO a candidate that deletes parts of the base in order to satisfy MAX-BR and the templatic requirements placed on the reduplicant is the most harmonic. Candidate (27a) presents this effect and the other candidates (27b, c) that violate either MAX-BR or the templatic constraint (RED = MinWd) are less harmonic.

There have been two types of responses to the Hamilton-Kager Conundrum. The first as outlined in McCarthy and Prince (1994*b*) is to deny or remove reduplicative "templates" from OT. This proposal is known as *generalized template theory* (GTT). The idea behind GTT is that if there are no constraints that state templatic requirements on reduplicants (such as RED = MinWd) then there is no way to create a constraint ranking like that in

(27) which will produce templatic backcopying. There are two main problems with this solution. The first is that it is not clear that all reduplicative templates can be produced through the interaction of other constraints as suggested by McCarthy and Prince. Kager (1999: 227) uses the constraint RED = σ and in a footnote states, "We leave open the consequences of adopting this constraint for Generalized Template theory." Kager's analysis highlights the uncertainty of the empirical adequacy of this aspect of GTT. Specific reduplication patterns that must be addressed are heavy syllables and VC(C) patterns because it is unclear how these prosodic shapes emerge from the interaction of prosodic well-formedness constraints. See Raimy (2000*a*) for further discussion of this issue.

The second problem with the GTT approach is that the generalizations about reduplication patterns that were encoded in templates disappear with the elimination of templates. One of the goals of generative phonology is to identify what are the relevant generalizations a learner has to make when acquiring a language. GTT denies that a speaker of a language makes any distinct generalization about a given reduplication pattern. Instead, constraints interact to produce the correct surface patterns of reduplication but this leads to a fractured generalization spread across numerous rankings between individual constraints. It is unclear whether this type of effect should be considered a generalization and it is even less clear that this type of generalization is an improvement over a template or adds to our knowledge about human language.

The other response to the Hamilton-Kager Conundrum is presented by Spaelti (1997). Spaelti argues that all 'reduplication specific' effects result from the *emergence of the unmarked* (McCarthy and Prince 1994*a*) ranking seen in (28).

(28) FAITH-IO >> Markedness >> FAITH-BR

This approach removes the Hamilton-Kager Conundrum not by eliminating specific constraints from Con but by limiting the possible constraint rankings. While the meta-ranking FAITH-IO >> FAITH-BR mimics the other stipulated meta-ranking FAITH-Root >> FAITH-Affix and thus appears to indicate that there is some generalization to be made about the structure of Con, the emergence of the unmarked solution to the Hamilton-Kager Conundrum also faces empirical problems. Specifically, FAITH-BR must be ranked higher than FAITH-IO in order to account for segmental backcopying effects. Consider the tableau in (29) which presents McCarthy and Prince's analysis of nasal spread in Malay (as presented by Kager 1999: 236).

(29)

	RED + waŋɪ	IDENT-BR (nasal)	*NV$_{ORAL}$	*V$_{NASAL}$	IDENT-IO (nasal)
☞	a. <u>w̃ãŋĩ</u>-w̃ãŋĩ			******	**
	b. <u>waŋĩ</u>-waŋĩ		*!*	**	
	c. <u>waŋĩ</u>-w̃ãŋĩ	*!*		****	**

The tableau in (29) shows that is it essential that IDENT-BR(nasal) is ranked higher than IDENT-IO(nasal) otherwise the backcopying of nasality in Malay can not be captured. Consequently we see that Spaelti's (1997) meta-ranking of FAITH-IO >> FAITH-BR is too restrictive and not empirically supported. One possible solution is to split FAITH-BR into MAX/DEP and IDENT families and only allow IDENT-BR to be ranked above FAITH-IO. This move is entirely stipulatory though and provides no explanation of the lack of Hamilton-Kager Conundrum effects.

More promisingly, the dynamic theory of the base allows a different solution to the Hamilton-Kager Conundrum. If we reconsider the tableau in (27), we can see that one of the crucial assumptions involved here is that the base is coextensive with the stem/root. Consider the tableau in (30) which abandons this strict assumption and allows the base to be determined through constraint interaction.

(30)

	/RED + tilparku/	RED = MinWd	MAX-BR	MAX$_{IO}$	BASE = STEM
	a. **tilpa**-<u>tilpa</u>			*!**	
	b. **tilpa**-<u>tilparku</u>		*!**		
	c. **tilparku**-<u>tilparku</u>	*!			
☞	d. **tilpa**-<u>tilpa</u>rku				***

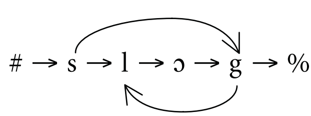
With the dynamic base constraint BASE = STEM ranked below MAX-IO, backcopying truncation no longer necessarily occurs when a reduplicative template and MAX-BR are ranked above MAX-IO. Whether the Hamilton-Kager Conundrum occurs or not is now a function of the ranking between FAITH-IO and whatever constraints determine the base. This is a welcomed result since backcopying of segmental processes can now be understood as emanating from the free ranking of FAITH-BR, FAITH-IO, and relevant markedness constraints. Templatic backcopying can be ruled out via a meta-ranking of FAITH-IO >> Base Constraints without affecting present analyses of segmental backcopying effects. While this solution to the Hamilton-Kager Conundrum still requires a stipulated meta-ranking, it does provide a principled answer as to why segmental processes and prosodic templates behave differently. The form of this explanation is that segmental backcopying only involves the interaction of markedness, FAITH-IO, and FAITH-BR, and backcopying is just one of the typological possibilities allowed by the free re-ranking of constraints. Templatic backcopying is not a typological possibility because of meta-ranking that restricts the family of Base Constraints to be ranked below FAITH-IO. The overall effect of this ranking is that when there is a conflict between high-ranking FAITH-BR and FAITH-IO, the response will be the modification of the base region in the output with no impact on the actual FAITH-IO mapping.

It can now be seen that the dynamic theory of the base in reduplication is the only empirically adequate and coherent hypothesis. The coherency of this hypothesis results from it utilizing the main mechanism of expressing generalizations available to OT, namely, constraint interaction. The empirical adequacy of this hypothesis resides in being able to capture reduplication patterns where the base and reduplicant are separated by segmental material that is not part of either in violation of the *adjacent string hypothesis* (ASH). We are not worried that ASH can be violated since one of the main tenets of OT is that all constraints can be violated. Cases where ASH must not be violated (as in Urbanczyk's work on Lushootseed) are accounted for by ranking ASH sufficiently high in the constraint hierarchy and cases where ASH is violated (as in Indonesian interposed reduplication) are handled with a lower ranking of ASH. Additional support for the dynamic base hypothesis is found in its novel solution to the Hamilton Kager Conundrum. By distinguishing between the interplay of the calculation of a base and prosodic requirements placed on the reduplication and the role IDENT plays along segmental dimensions we can begin to see why prosodic templates never backcopy. All of these points indicate that within the OT research program the idea of a dynamic base in reduplication should be pursued further.

* *

Having now determined that the dynamic base hypothesis appears to be the most adequate way of understanding aspects of the GenOUTPUT structure produced in reduplicated structures, we can return to the main issue of this section. This issue is acquisition and how the structure of Gen affects it. For reasons of clarity and space, only GenOUTPUTS that coincide with the targeted output form for the Temiar example will be considered in this discussion. Each of the different approaches to calculating the base and reduplicant allow for a plethora of less harmonic GenOUTPUTS to be possible candidates. Space precludes us from presenting all of these failed candidates but a learner does not have the luxury of limiting their hypothesis space like this. The fact remains that all of these alternatives must be considered by the child who is learning Temiar. This is the main problem with the Correspondence Theory of reduplication. We see above that when the positions on reduplicative structures are culled from the present literature and one that is theoretically coherent and empirically adequate is settled on, there are vast possibilities of GenOUTPUTS that must be considered when a child is acquiring a language. To compound this problem it is not clear which of the GenOUTPUT structures discussed in this section for Temiar is the most harmonic. Each one can be the most harmonic based on some ranking of constraints which leads to the question of which grammar the child should arrive at. If our theory does not provide a way of determining what grammar we think a child should be striving for, then we will never be able to explain how language is acquired.

The Raimy model of reduplication does not suffer from this analytic indeterminacy. The representation the child should acquire for the Temiar form in (4) is the representation in (5) repeated below as (31).

(31) 

The representation in (31) is the simplest representation possible within the Raimy model of reduplication that will produce the correct surface form. Any further addition of precedence links to (31) only complicates the representation further without producing any benefit in computation or empirical adequacy. The fact that there is a single representation that can be easily identified through the metric of analytic simplicity indicates that the Raimy model provides a more constrained hypothesis space to the learner. There is no analog to the question of what the correct GenOUTPUT structure is for Temiar for the Raimy model. This produces a strong argument in favor of the Raimy model of reduplication over the Correspondence Model. Since the Raimy model provides a more constrained hypothesis space to a learner for what the representation of a reduplicated form should be, this model of reduplication should be preferred over the Correspondence Model of reduplication. To further support this conclusion, it must be recognized that the hypothesis space for the child acquiring an OT grammar is actually much worse than presented in this section if full and free constraint re-ranking is possible and if Eval considers every possible candidate produced by Gen.

5.8 Computational profligacy in Correspondence Theory

The above discussion of GenOUTPUT structures shows the analytic indeterminacy present in CT with respect to analyzing reduplication. This indeterminacy directly results from the massive expressive power of the system based on the freely generated non-morphological and non-prosodic output structure of RED. This amount of expressive power also has implications for language typology. As mentioned in Section 5.2, the Raimy model has a natural markedness metric based on formal complexity but CT does not. The lack of this markedness metric in CT prevents useful typologies from being constructed with presently accepted constraints. Consider the data in (32) which shows perfective reduplication of full grade stems in Sanskrit (data is taken from Kager (1999) who cites Steriade (1988) and Whitney (1889)).

(32) pa-prat^{h}-a 'spread'
ma-mna:-u 'note'
sa-swar 'sound'
da-d^{h}wans-a 'scatter'

Kager (1999: 214–15) discusses the relevance of this reduplication pattern as an instance of the emergence of the unmarked. Specifically, Kager presents the constraint ranking in (33) to account for this pattern.

(33) Faithfulness >> Well-Formedness >> Reduplicative Identity
MAX-IO >> *COMPLEX >> MAX-BR, CONTIGUITY-BR

The ranking in (33) appears to be innocuous and is required in order to account for the Sanskrit data in (32) but does not tell the whole story of this reduplication pattern. One aspect omitted from Kager's analysis of this pattern is how the reduplicant is limited in shape to a single syllable. This omission does not alter Kager's main point of discussing this reduplication pattern that contiguity can be violated in base-reduplicant mappings but it does leave open what ramifications this fact has for typological claims made by OT. What shape restriction is put on the reduplicant is crucially important in fully accounting for reduplication in Sanskrit. Consider the tableau in (34). Since there are no "shape" or "restrictor" constraints ranked above MAX-BR in this tableau, total reduplication will result. Note that we will assume that "the base is the stem" for expository reasons and this facet of the analysis can be changed if need be.

(34)

	RED + prat^h-a	MAX-IO	*COMPLEX	MAX-BR	CONT-BR
	a. pa-prat^{h}a		*	**!*	*
	b. pra-prat^h-a		**!	**	
	c. pat^h-prat^h-a		*	**!	*
☞	d. pat^{h}a-prat^{h}a		*	*	*

The most harmonic candidate in (34) is (34d) which copies the entire base except for /r/ which would create a complex onset in the reduplicant.[10] All other candidates either copy too little of the base (34a, c) or copy the complex onset (34b). Kager's sketch of Sanskrit can be saved by simply ranking RED = σ "Align both edges of the reduplicant with the edges of a syllable" (Kager 1999: 227) and NOCODA above MAX-BR. This produces the tableau and results in (35).

(35)

	RED + prat^h-a	MAX-IO	*COMPLEX	RED = σ	NOCODA	MAX-BR	CONT-BR
☞	a. pa-prat^{h}a		*			***	*
	b. pra-prat^h-a		**!			**	
	c. pat^h-prat^h-a		*		*!	**	*
	d. pat^{h}a-prat^{h}a		*	*!		*	*

[10] We assume that DEP-IO and/or DEP-BR is ranked above *COMPLEX to prevent a candidate such as **pirat^{h}a-prat^{h}a* from emerging as optimal.

(35) shows the necessity of ranking some sort of "shape" constraints above Max-BR in order to derive the occurring surface forms in Sanskrit perfective reduplication. From this point we can now ask a typological question. Specifically, what are the typological predictions given the constraints used in (35) under different rankings? More to the point, consider the particular ranking presented in tableau (36) for a hypothetical input *prabtru*. What does this tableau tell us?

(36)

	RED + prabtru	Max-IO	*Complex	Max-BR	Cont-BR	Red = σ	NoCoda
	a. pa-prabtru		**	***!**	*		*
	b. pra-prabtru		***!	****			*
	c. pab-prabtru		**	***!*	*		**
☞	d. pabtu-prabtru		**	**	**	*	**
	e. prabtru-prabtru		***!*			*	**

Candidate (36d) is the most harmonic candidate given the ranking in (36). This particular candidate copies all segments except for ones that would lead to violations of *Complex. Other candidates fare less well since they either copy less of the base (36a, c) or have gratuitous violations of *Complex (36b, e).

The results found in (36) appear to be the standard state of affairs within OT until we recognize the fact that the constraint ranking in (36) characterizes a non-attested pattern of reduplication. The ranking in (36) produces a reduplication pattern that will simplify every complex onset in a reduplicant regardless of how many there are. There is no attested reduplication pattern where total reduplication occurs except for the deletion of segments in complex onsets and this is the pattern that the constraint ranking in (36) characterizes. The important aspect to see here is that there is no apparent way to distinguish the constraint ranking in (35) which produces an attested natural human language and the constraint ranking in (36) which produces a pattern that is not found in natural human language.

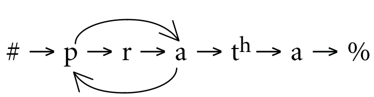
The free re-ranking of constraints to produce typologies is not a beneficial feature of OT grammars. In fact, results like those found in (36) indicate that the typologies created by the free re-ranking of constraints are a liability to OT since it is as easy to produce unattested patterns as it is to produce attested patterns. Not being able to distinguish between occurring and non-occurring patterns suggests that the CT model of reduplication is too powerful to produce a constrained hypothesis space that could guide acquisition.

The results in (36) are not unique within Correspondence Theory since Stemberger (1996)[11] points out that analogous non-occurring reduplication patterns can be produced by ranking NoCoda above Max-BR and Contiguity-BR. Consider the tableau in (37) which shows this result for the hypothetical input *pabtup*.

(37)

	RED + pabtup	Max-IO	NoCoda	Max-BR	Cont-BR
	a. pa-pabtup		**	***!*	
	b. pab-pabtup		***!	***	
	c. pabtu-pabtup		***!	*	
	d. pabtup-pabtup		***!*		
☞	e. patu-pabtup		**	**	

The common theme that emerges when we consider tableaus (36–7) is that there is no necessary connection between total reduplication and maintaining reduplicant internal contiguity in the OT approach to reduplication. In contrast to this, natural human language appears to connect total reduplication with respecting the contiguity of the base. The Raimy model of reduplication has this characteristic. To see how this result obtains, we will begin by seeing how the Sanskrit reduplication pattern in (32) is accounted for in the Raimy model. Consider the representation in (38).

(38) 

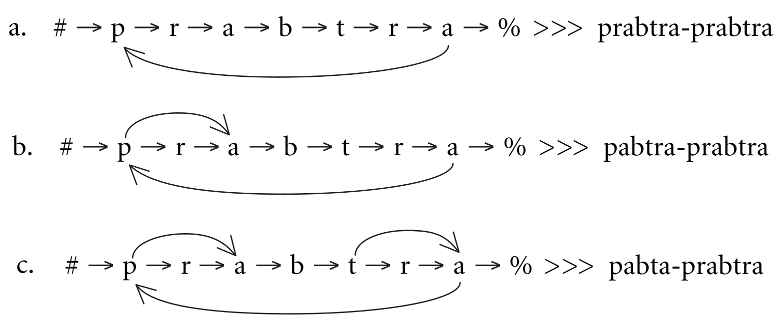
>>> pa-pratʰa

[11] Stemberger (1996) contains analogous arguments based on the *Complex Onset facts which inspired the discussion of the Sanskrit facts in this chapter.

(38) presents the precedence graph that is required to account for the Sanskrit 'core syllable' reduplication pattern presented in (32). The important thing to recognize about the graph in (38) is that two distinct links, [p → a] and [a → p], must be added to produce this reduplication pattern. The link [a → p] creates the loop that causes repetition of segmental material. If only this link is added, simple light syllable reduplication (*prat*^{h}*a* >> *pra-prat*^{h}*a*) is produced. The additional link from [p → a] causes the surface appearance of the reduplicant violating the contiguity of the base. This jump link must be independently specified as distinct from the reduplicative back link. The markedness of this type of reduplication pattern is immediately captured within the Raimy model since two distinct precedence links must be added. Less marked reduplication patterns (ones where the contiguity of the base and reduplicant do not diverge) only require a single precedence link to be added. This is a desired attribute since the metric of analytic simplicity allows marked and unmarked reduplication patterns to be easily distinguished within the Raimy model.

With this result in hand, we can now investigate how contiguity in the base and reduplicant is preserved within the Raimy model. Consider the representations in (39).

(39)



(39a) shows a precedence graph that will result in total reduplication. There is no possible way for the reduplicant to diverge from the base with respect to contiguity since there is only a single precedence path through the base in this representation. (39b) presents a graph where an additional link has been added which creates an alternative precedence path through the base. Since there are now two distinct paths through the base, the base and reduplicant can diverge along the dimension of contiguity. As pointed out in Raimy (2000*a*) the additional link from [p → a] will be followed first given the nature of the linearization process, thus the surface form of (39b) indicates a prefixing pattern of reduplication.

(39c) presents a precedence graph that coincides with a reduplicant that strips out every complex onset that occurs in a base. This representation is equivalent to the output form produced by the constraint ranking in (36). Immediately, we can see that a difference between (39b) and (39c) is that (39c) has added another jump link in order to omit this complex onset from the reduplicant in the linearized form. We can generalize this behavior in that an extra jump link needs to be added for each complex onset that needs to be omitted in the output. The dependency between characteristics of the base and the number of links added is one that does not appear to occur in natural human language. This provides an explanation for why contiguity violations do not occur in patterns of total reduplication.

The Raimy model of reduplication derives this behavior of contiguity preservation in total reduplication patterns from the fundamental principles on how reduplicative structures are built. Total reduplication results from the addition of a precedence link from the end of a form to the beginning of the form. This additional link does not alter the precedence path through the base in any way. In order to produce the effect of omitting all complex onsets (or codas) in the reduplicant a variable number of additional links must be added to the precedence graph that is dependent on how many complex onsets there are present in the precedence graph.

The only obvious way of producing the complex onset-stripping behavior in the reduplicant is to make the rule that adds the needed jump link iterative.[12] This is an entirely *ad hoc* move with no motivation behind it. There is no reason why the type of morphological rule involved in adding a jump link should be iterative. Proposals on iterative rules (Myers 1991; Halle and Vergnaud 1987) limit iteration to phonological rules that define either inventories or well-formedness aspects of representations. Morphological rules that add precedence links do not fall under either of these categories. Since there is no way to motivate the iteration of the jump link rule, there is no way (other than a brute force stipulation) for the Raimy model to produce a pattern of total reduplication with word-internal contiguity violations. This is another welcomed result since this limit in the productive power of the Raimy model coincides with what we know about the existing patterns of reduplication in the world's languages.[13]

[12] Interestingly, ludlings may differ precisely on this point (Bagemihl 1995).

[13] Given this restriction on what an iterative rule in the phonology is, we can hypothesize that iterative morphological rules are ludlings.

A final argument which indicates that the CT model of reduplication dramatically overgenerates possible reduplication patterns is based on string reversal. Pinker and Prince (1988) discuss the relevance of string reversal in evaluating representational models because:

> [the] most challenging requirement we can place on a representational system is that it should exclude the impossible. Many kinds of formally simple relations are absent from natural language, presumably because they can not be mentally represented . . . A quintessential unlinguistic map is relating a string to its mirror image reversal (this would relate *pit* to *tip*, *brag* to *garb*, *dumb* to *mud*, and so on); although neither physiology nor physics forbids it, no language uses such a pattern.
>
> (Pinker and Prince 1988: 99–100)

Stemberger (1996) argues that CT easily produces string-reversal reduplication patterns. Consider the tableau in (40) (taken from Stemberger 1996: 148).

(40)

	/akison/	ONSET	CONTIGUITY	LINEARITY	ANCHORING
	a. akison	*!			
☞	b. nosika			*...	*...
	c. kasino		*!	*...	*...
	d. nakiso		*!	*...	*...

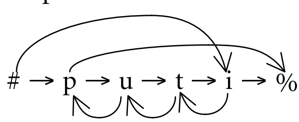
The constraints in (40) are all well-accepted ones (all proposed in McCarthy and Prince 1995*a*). What Stemberger has noticed is that if ONSET and CONTIGUITY are ranked above LINEARITY and ANCHORING with MAX and DEP ranked above ONSET (this is omitted in the tableau) then vowel-initial inputs will string-reverse. Candidate (40b) is the most harmonic since there are no violations of ONSET and CONTIGUITY. If free re-ranking of constraints is the underlying principle or basis of typologies within CT (or Optimality Theory in general) then we must conclude that CT is much too powerful a model of grammar.

The problem that arises within CT is not that string reversal can be produced but that there is no method of distinguishing between a ranking of constraints that produces string reversal and constraint rankings that bar string reversal. An adequate model of grammar should be able to characterize this distinction in some manner. The Raimy model of reduplication has a natural way of making this distinction. First we must note that string reversal is found in some language games in natural language. Bagemihl (1995) offers the example from Tagalog presented in (41).

(41) *Golagat* (Gil 1990)
puti 'white' > itup

Further evidence for the possibility of string reversal in human language is found in the enjoyment of palindromes. The issue here is how to characterize processes that only occur in language games versus processes that occur as part of a grammar of a natural human language. If we consider how string reversal is accomplished in the Raimy model, a natural solution to this dilemma is seen.

(42) a. # → p → u → t → i → %

b. 

>>> itup

c. (i) ADD # → [_%] "make last segment the first segment"
(ii) ADD [#_] → % "make the first segment the last segment"
(iii) ADD A → [B /_→ A] *iterate*
"add a link from a following segment to the preceding segment and iterate"

(42b) shows the precedence graph that is required in order to produce the surface effect of string reversal and (42c) presents the algorithm required to produce this precedence graph. There are at least two ways in which the algorithm in (42c) which characterizes string reversal in the Raimy model is crucially distinct from processes found in natural language. The first, which has already been discussed in reference to complex onset simplification in total reduplication patterns, is that to produce string reversal in the Raimy model an iterative process must be invoked as indicated in (42c, iii). The second way string reversal is distinct from natural processes is in the number of precedence links that must be added. In addition to the iterative process of adding links in (42c, iii) two additional distinct precedence links must be added (42c, i/ii). This results is a total of three distinct link adding components to string reversal with one of these processes necessarily being iterative. These characteristics clearly indicate that string reversal is a much more complicated process than reduplication or simple affixation. This is the exact result that we want from a representational system and it highlights further the distinction between CT and the Raimy model of reduplication.

To summarize this section, both CT and the Raimy model of reduplication provide adequate empirical coverage of the known patterns in reduplication. Despite this equivalence, CT and the Raimy model differ in whether the computation of "natural" patterns found in human language is distinguished from the computation of "unnatural" patterns. CT is such a powerful computational system that it is unable to distinguish between natural and unnatural patterns in human language. In other words, CT predicts that unnatural patterns are as likely to occur as natural patterns are in human language. The Raimy model on the other hand easily distinguishes between natural and unnatural patterns based on the complexity of the required computation to produce the required precedence graphs. Simple computations such as the addition of a single link in order to produce reduplication map onto unmarked human language processes. Slightly more complex computations such as the addition of two links to account for the Sanskrit complex onset elimination pattern of partial reduplication are also available in natural human language but they are "marked." Finally, truly complex operations only appear in the realm of language games where the linguistic grammar is utilized beyond its normal limits in a creative manner.

5.9 Summary

Both theories of reduplication discussed here require new representational resources–OT adds Correspondence Theory, Generative Phonology adds non-linear temporal relations proposed by Raimy (2000*a*).[14] Correspondence Theory has been extremely useful in analyses of reduplication primarily because it is so powerful. In fact, it is overly powerful in representational possibilities in that it induces an exponential explosion in the number of candidates which are distinct phonologically but identical phonetically. On the computational side, CT is also overly powerful since the constraints added to Con when freely re-ranked to create language typologies produce non-occurring patterns of language as easily as they produce occurring patterns. In contrast, the Raimy model's introduction of non-linear temporal links to phonological representations adds the minimum amount of power necessary to describe the concept of repetition. Non-linear links have several other advantages, such as deriving markedness relations, capturing modularity considerations, and

[14] One could also pursue the other logical possibilities, OT with non-linear temporal relations or Generative Phonology with correspondence.

making predictions about the behaviour of other potentially similar phenomena (geminate integrity, truncation effects, etc.).

In conclusion, since OT and Generative Phonology now have equivalent (or nearly equivalent) empirical coverage with respect to reduplication, two final points should be made. The first point is that argumentation based on empirical facts from reduplication no longer directly bears on the issue of computational differences between OT and Generative Phonology. McCarthy and Prince (1995*a*) first presented this type of argument by showing that serial models of reduplication are incapable of capturing backcopying phenomena. Raimy (2000*a*, *b*) addresses these arguments and illustrates how the enhanced precedence representations provide a conceptually elegant and empirically adequate analysis of backcopying effects. The second and main point of this chapter is that since the Raimy model of reduplication presents a less powerful change to representations in phonology than Correspondence Theory does, it should be adopted as a more desirable and explanatory theory of representations and reduplication. Because the Raimy model of reduplication is less powerful than CT, it offers a more constrained grammar space which provides a more tractable learnability problem for children. Representations for reduplication in the Raimy model are easier to learn because there is only a single possible representation for a given reduplication pattern, thus allowing a cue/trigger-based learning algorithm that simply notices repetitions in a string of phonemes. The CT model of reduplication not only has to provide a method for the learner to notice reduplication but also has to indicate how the learner chooses between possible representations of a given reduplication pattern. The fact that there is no choice among different representations for specific reduplication patterns in the Raimy model argues strongly that this model is more constrained than CT which provides multiple possible representations for any reduplication pattern. Since the Raimy model is more constrained than CT it provides more explanatory analyses of reduplication and thus should be preferred.

6

Fenno-Swedish Quantity: Contrast in Stratal OT

PAUL KIPARSKY

6.1 The weight of stressed syllables

Compared to more familiar varieties of Swedish, the dialects spoken in Finland have rather diverse syllable structures.[1] The distribution of distinctive syllable weight is determined by grammatical factors, and by varying effects of final consonant weightlessness. In turn it constrains several gemination processes which create derived super heavy syllables, in an unexpected way which provides evidence for an anti-neutralization constraint. Stratal OT, which integrates OT with Lexical Phonology, sheds light on these complex quantity systems.

6.1.1 *Light stressed syllables*

The bimoraic minimum: Sweden vs. Finland In most Swedish dialects of Sweden (here referred to as *West Swedish* for short), stressed syllables are minimally bimoraic: they must contain at least a long vowel (-VV-) or a closed syllable (-VC-).[2] Words like (1a) are therefore impossible. Because word-final consonants are weightless ("extrametrical") in Swedish, the two-mora minimum also excludes monosyllabic words with -VC rhymes (see (1b)):

(1) a. *[ro], *[ro.da], *[ro.a], *[no.gra]

b. *[ro(d)]

[1] The information on Fenno-Swedish dialects given here is based primarily on the 29 transcribed dialect texts in Harling-Kranck 1998, with accompanying tapes, as well as on the brief grammatical sketches of the dialects provided there. Page references below are to that work, unless otherwise specified. For supplementary information on particular points I have consulted the additional dialect monographs cited below. Special thanks are due to Mikael Reuter, for valuable discussion of Helsinki Swedish, and for generously providing me with a copy of his unpublished thesis (Reuter 1982).

[2] Except where otherwise stated, the generalizations stated here hold for phonological words. Each member of a compound constitutes a separate phonological word.

The Swedish dialects of Finland present a more varied picture. Only parts of Åland have the two-mora minimum (e.g. Brändö and Kumlinge in the northeastern part of the island, Sundberg 1993: 131 ff.). All other Fenno-Swedish dialects allow light (i.e. monomoraic) stressed syllables as a distinctive syllable type:

(2) [daga], [dagar] 'days', [viku] 'week', [veliŋ] 'gruel', [suvel] 'food eaten with bread, sowl', [somar] 'summer', [stygu] 'hut', [päron] 'potato(es)', [hakon] 'the chin', [hole] 'the hole', [segla] 'to sail', [tala] 'to talk', [sita] 'to sit', [myky][3] 'much', [staduɡari] 'steadier', [snidit] 'askew', [ʧyvu] 'twenty'

Fenno-Swedish, then, has a lexical contrast between stressed CV, CVC, and CVV syllables:[4]

(3) a. [baka] 'bake' (99), [baaket] 'after' (adv.) (114), [bakkan] 'the hill' (114)
b. [vaten] 'water' (102), [maaten] 'the food', [natten] 'the night' (70)
c. [betär] 'better' (51), [fleetor] 'braids' (43), [tvettar] 'washes' (51)

Even though stressed CV *syllables* are allowed, *words* of the form CV are categorically excluded in all the dialects (except for function words, on which see below).[5] As for words of the form CVC, the dialects are divided. Most allow them:[6]

(4) [sov] 'slept' (21), [styd] 'support' (22), [hol] 'hole' (22), [led] 'opening (in fence)' (31), [smör] 'butter' (55), [lag] 'to make' (55), [rog] 'rye' (134), [tär] 'there' (129), [las] 'read' (past) (Huldén 1957: 133), [far] 'rides' (Huldén 1957: 165), [net] 'net', [skot] 'shot', [gres] 'grass' (Selenius 1972: 34)

CVC words are excluded, however, in southern Ostrobothnia, on some islands off Turku/Åbo in the southwest, and, as already mentioned, in the

[3] Here and throughout I ignore dialectal variation in pronunciation where it is not relevant to the analysis of syllable weight. For instance, dialects with palatalization before front vowels have [myʧy] or [myʧi] instead of [myky].

[4] In phonetic transcriptions of Fenno-Swedish, I adhere to IPA standards except that I mark vowel and consonant length by gemination, which makes it easier to visualize syllable weight, and allows convenient marking of syllable boundaries (by "."). Italics are reserved for citing words in Swedish spelling, which will be done for standard West Swedish and standard Helsinki Swedish only; in those transcriptions I supply macrons to mark vowel length.

[5] The single contrary example is *ga* [ga] 'go' in Vörå (central Ostrobothnia, Harling-Kranck 1998: 121), apparently a fast speech variant of that dialect's normal [gaa].

[6] The contrast between /CVC/ and /CVCC/ is clearest before a vowel in close contact, e.g. [hol i mitten] 'hole in the middle', [rʉnn o] 'round too' (Harling-Kranck 1998: 22). The /CVC/ words are partly retentions of Proto-Nordic /CVC/, partly analogical reintroductions (Huldén 1957: 122), and partly apocopated from CVCV at different periods.

Åland dialects that impose the West Swedish two-mora minimum on stressed syllables.

In the urban Fenno-Swedish of Helsinki and Turku, light stressed syllables have a more restricted distribution. Open syllables in lexical words (such as the nouns, verbs, adjectives, and adverbs in (2)) are obligatorily lengthened under stress, as in Sweden. Light stressed syllables do occur, but only in certain rather special circumstances: in function words before voiced consonants, in truncated lexical words (such as (5f)), and a few others discussed below. Consequently, Helsinki/Turku Swedish does not have the particular three-way contrasts in (3), though it still has those in (5).[7]

(5) Helsinki/Turku Swedish:

a. *före* [före] 'before', *före* [fööre] 'ski trail conditions', *förre* [förre] 'former'
b. *bara* [bara] 'only', *bara* [baara] 'the bare', *barra* [barra] 'to shed needles'
c. *hela* [hela] 'the whole' (all of), *hela* [heela] 'the whole' (undamaged), *hälla* [hella] 'to pour'
d. *mina* [mina] 'my', *mina* [miina] 'mine' (explosive device), *minna(s)* [minna(s)] 'to remember'
e. *så* [so] 'so', *så* [soo] 'to sow'
f. *dia* [dia] 'slide, transparency', *dia* [diia] 'to suckle'

The core constraints The data so far have a fairly straightforward analysis, except for the mysterious restrictions in Helsinki/Turku, to which I return below after surveying the other parameters of syllable weight. Let us assume the constraints in (6):

(6) a. CONSONANT EXTRAMETRICALITY (abbreviated C-Ex): A word-final consonant is weightless (i.e. it is not part of the prosodic word).
b. FOOT-BINARITY: A foot (and hence a word) has at least two moras.
c. STRESS-TO-WEIGHT: A stressed syllable has at least two moras.
d. DEP-Vμ: An output vocalic mora corresponds to an input mora ("don't lengthen vowels").

[7] The Helsinki/Turku data, and most of the descriptive generalizations discussed below, are from Reuter (1982) (especially valuable for its phonetic data), Reuter (1986), and Bergroth (1928). This variety of Swedish is essentially identical with the one I learned in Helsinki in the 1940s and early 1950s.

The most widespread type of Fenno-Swedish, where /CVC/ words remain unlengthened, is derived by the following ranking (where commas separate constraints whose mutual ranking is not crucial):

(7) General Fenno-Swedish: Foot-Binarity ≫ Dep-Vμ ≫ Stress-to-Weight, C-Extrametricality

(8)

General F.-Sw.		Ft-Bin	Dep-Vμ	Str/Wt	C-Ex
Input: /CVCV/					
1a. ☞	CV́.CV			*	
1b.	CV́V.CV		*		
Input: /CVC/					
2a.	CV́(C)	*		*	
2b. ☞	CV́C				*
2c.	CV́V(C)		*		
Input: /CV/					
3a.	CV́	*		*	
3b. ☞	CV́V		*		

Its similarity to Proto-Nordic, and its discontinuous distribution within Finland, suggest that this is the most archaic of the Fenno-Swedish quantity systems. Pointing to the same conclusion is the formal relationship between the constraint systems of the dialects. In the Stratal OT framework, (Booij 1996, 1997; Orgun 1996*d*; Kiparsky 2000, 2003; Bermúdez-Otero 1999, 2006*c*, 2006*d*; Bermúdez-Otero and Hogg 2003; Rubach 1997, 2000) sound change corresponds the promotion of markedness constraints to undominated status in the postlexical phonology (with the innovative constraint ranking then spreading to the word phonology, or even to the stem phonology). If (7) is taken as the point of origin, each of the attested systems is derivable from another by a single constraint promotion.

Starting from (7), promotion of Consonant Extrametricality to undominated status yields the ranking in (9), which characterizes the dialects of South and Central Ostrobothnia and of the southwestern islands:

(9) South Ostrobothnia: Consonant Extrametricality, Foot-Binarity ≫ Dep-Vμ ≫ Stress-to-Weight

In these dialects, input words of the form /CVCV/, /CVC/, and /CV/ surface respectively as [CV.CV], [CVVC], and [CVV].

(10)

S. Ostrobothnia	C-Ex	Ft-Bin	Dep-Vμ	Str/Wt
Input: /CVCV/				
1a. ☞ CV́.CV				*
1b. CV́V.CV			*	
Input: /CVC/				
2a. CV́(C)		*		*
2b. CV́C	*			
2c. ☞ CV́V(C)			*	
Input: /CV/				
3a. CV́		*	*	
3b. ☞ CV́V			*	

As the constraints correctly predict, lengthened monosyllabic bases alternate with short-vowel suffixed forms in South Ostrobothnian:[8]

(11)	[faar] 'rides'	[fara]	'to ride'	(Nagu, 153)	
	[veed] 'wood'	[vedin]	'the wood'	(Lappfjärd, 99)	
	[koom] 'came'	[koma]	'to come'	(Petalax, 109; Munsala, Huldén 1957: 125)	
	[taal] 'speech'	[talar]	'speaks'	(Närpes, Riad 1992: 181)	
	[viik] 'week'	[vikun]	'the week'	(Närpes, Riad 1992: 181)	

If, in addition, Stress-to-Weight is promoted, we get the dialects with consistent open syllable lengthening, such as the Swedish of Åland and Sweden:[9]

(12) West Swedish: Stress-to-Weight, C-Extrametricality, Foot-Binarity ≫ Dep-Vμ

The input words /CVCV/, /CVC/, and /CV/ then surface respectively as [CV.CV], [CVVC], and /CVV/:

[8] Analogous length alternations have developed in the dialect of Älvdalen in Sweden, e.g. *smiið* 'blacksmith', pl. *smiðir*; *daal* 'valley', pl. *dalir* (Riad 1992: 306). They are also found (but before final obstruents only) in the German dialects of northeastern Switzerland, e.g. *šmiid* 'smith', *šmidə* 'to forge'; *baad* 'bath', pl. *bedər*; *glaas* 'glass', pl. *glesər* (Toggenburg, Wiget 1916: 70; Glarus, Streiff 1915: 49; Thurgau, Kraehenmann 2001*a*, 2001*b*).

[9] As well as, of course, of Icelandic (Kiparsky 1984).

(13)

West Swedish		Str/Wt	C-Ex	Ft-Bin	Dep-Vμ
Input: /CVCV/					
1a.	CV́.CV	*			
1b. ☞	CV́V.CV				*
Input: /CVC/					
2a.	CV́(C)	*		*	
2b.	CV́C		*		
2c. ☞	CV́V(C)				*
Input: /CV/					
3a.	CV́	*		*	
3b. ☞	CV́V				*

A fourth system emerges if at stage (7) Stress-to-Weight (rather than C-Extrametricality) is promoted. This is the standard Danish system, with open syllable lengthening but no monosyllable lengthening:[10]

(14) Danish: Stress-to-Weight, Foot-Binarity ≫ Dep-Vμ ≫ C-Extrametricality

(15) [glad] 'happy' [glaade] 'happy' (pl.)
[blad] 'leaf' [blaadet] 'the leaf'
[gud] 'god' [guuden] 'the god' (Riad 1992: 330)

The reader can verify that if some other system than (7) were instead posited as the original one, then (7) and (9) could not be derived from it by constraint promotion without positing unattested intermediate stages.[11] This confirms that the dialect with (7) is the most conservative.

6.1.2 *Distinctive superheavy syllables*

The behavior of superheavy syllables is clear-cut in the special case when they contain a long vowel followed by a geminate consonant. In West Swedish, these are categorically excluded in stems, and stem-final long vowels are shortened before suffixes beginning with geminates.

[10] The same alternation is found in noun inflection in certain north German dialects, e.g. *Glas* 'glass', pl. *Gläser*; *Rad* 'wheel', gen. *Raades* (only before final obstruents); also in Dutch nouns, e.g. *dag*; pl. *daagen* 'day'; *glas*, pl. *glaazen* 'glass'; *hol*, pl. *hoolen* 'hole'.

[11] The West Swedish system (12) could in principle have arisen by the same two sound changes in reverse order. Perhaps this is what in fact happened in the Danish-type dialects of southern Sweden.

(16) *rodde* /ruu-dde/ [rudde] 'rowed' (cf. *ro* /ruu/ [ruu] 'row')

Because final -C is weightless, CVVCC words pattern with medial CVVC syllables:

(17) *rodd* /ruu-dd/ [rudd] 'rowing', *rott* /ruu-tt/ [rutt] 'rowed'

Outside of such gemination cases, stressed -VVC and -VCC syllables do occur in West Swedish, as do monosyllabic words in -VVCC and -VCCC. Contrast (18a) and (18b).

(18) a. *vikta* /viik-t-a/ [viikta] 'folded' (pl.), *vikt* [viikt] (sg.) (from *vika* [viikka] 'to fold')
b. *vikt-a* /vikt-a/ [vikkta] 'to weight' (e.g. in the statistical sense, from *vikt* [vikkt] 'weight')

In fact, all varieties of Swedish seem to have them, albeit with many phonological and morphological restrictions.[12]

The fact that the long vowel + geminate configuration is specially restricted can be explained on the basis of moraic theory as follows (Riad 1992: 244). If vowel length and consonant gemination are represented moraically, then a long vowel must correspond to two moras, and the first half of a geminate consonant must correspond to a mora.[13] Therefore a -VVC rhyme whose final -C initiates a geminate must contain *three* moras. Other kinds of -VVC rhymes *can* be trimoraic, but need not be, for rhyme consonants need not be weight-bearing—an analytic option not available when the -C is part of a geminate. Thus the modern Swedish dialects support Riad's (1992: 244) argument from earlier stages of Swedish for the intrinsically trimoraic character of the long vowel + geminate configuration (what he calls "true overlength"). In what follows I take this special type of -VVC syllable as a diagnostic of a dialect's superheavy syllables, on the assumption that other kinds of -VVC syllables are not necessarily superheavy (though they may be if the facts so dictate).

With respect to such intrinsic superheavy syllables, Fenno-Swedish dialects are again more permissive than those of Sweden. The dialects of Nyland (Uusimaa) and of northern and central Ostrobothnia allow them:

[12] For example, long vowels are generally allowed before obstruent + sonorant clusters, even if they are not possible onsets, e.g. *odla* [uud.la] 'cultivate', *tävla* [teev.la] 'compete'. On the other hand, *[uul.va], *[teel.va] are not possible Swedish words.

[13] On the treatment of initial geminates, as moraic semisyllables, see Kiparsky (2003).

(19) /loo-dde-s/ [looddes] 'pretended' (66); /dreett-en/ [dreetten] 'the shaft' (43)

Due to the weightlessness of final -C, these same dialects also have monosyllabic words of the form CVVCC, where CC is a geminate, as in (20) (contrast (17)).

(20) /smoo-tt/ [smoott] 'little one' (21) (from [smoo] 'little'); /haa-dd/ [haadd] 'had' (157) (from [haa] 'have'); /ruu-dd/ [ruudd] 'rowing' (from [ruu] 'row')

In these dialects, the shortening process seen in (16) and (17) simply does not apply. Superheavy syllables are lexically distinctive and contrast on the surface with the other three syllable types in (3). The same four-way contrast CVC : CVVC : CVCC : CVVCC is also found in monosyllabic words before *-t, -d, -s*:

(21) a. [led] 'opening (in fence)' (31); [(far-)leed] '(shipping) channel' (Selenius 1972: 210); [redd] 'afraid' (34); [beedd] 'asked' (pp.) (Huldén 1957: 146)

b. [skot] 'shot' (Selenius 1972: 34); [boot] 'boat' (Selenius 1972: 210); [pott(-stuul)] 'potty(-chair)' (22); [goott] 'gone' (39)

In the phonology of these dialects, the faithfulness constraint (22a) MAX-μ outranks and defeats the constraint (22b) $^*\mu\mu\mu$, which imposes the two-mora maximum on syllables.

(22) a. MAX-μ: An input mora corresponds to an output mora ("don't shorten syllables").

b. $^*\mu\mu\mu$: No three-mora syllables (Kager 1999).

Superheavy syllables also respond to final consonant weightlessness, but in a different way than monomoraic syllables do. Suppose that prosodic repair is prevented by high-ranking MAX and DEP constraints. Then, if the constraint requiring final -C to be weightless outranks prosodic minimality conditions (such as the requirement that feet have at least two moras), it prevents words that would otherwise satisfy them from doing so. -C weightlessness also *allows* the satisfaction of prosodic *maximality* conditions (such as the requirement that feet have at most three moras) by words that would otherwise violate them. But this second effect is *not* dependent on the mutual ranking of the constraints in question. Only the *prohibition* of C-Extrametricality could "bleed" a maximality constraint. Suppose there are no constraints that prohibit C-Extrametricality. Then an extra word-final consonant would be allowed on top of the three-mora syllable maximum in *all* dialects, and indeed the same should be true for all maximality conditions in all languages. It remains to be seen if this simple and strong hypothesis can be maintained.

A further argument for the moraic analysis of geminates comes from the consonant lengthening processes of Fenno-Swedish examined in the next subsection.

6.1.3 *Gemination and redundant superheaviness*

Coda Gemination Most Swedish dialects (possibly all of them) lengthen coda consonants after short stressed vowels. For the reasons stated below, the lengthened consonants will be considered true geminates.

(23) Coda gemination:

a. *vissna* [viss.na] 'to wilt', *vända* [venn.da] 'to turn', *stövlar* [stövv.lar] 'boots', *halva* [hall.va] 'half' (def.), *aska* [ass.ka] 'ash', *taxa* [takk.sa] 'rate'

b. *vikt* [vikkt] 'weight', *kraft* [krafft] 'strength', *visst* [visst] 'certainly', *vänd* [vennd] 'turn!', *golv* [gollv] 'floor', *bild* [billd] 'picture', *hund* [hʉnnd] 'dog'

In one special environment, most Fenno-Swedish dialects lengthen not the postvocalic coda consonant but the consonant after it, namely when the postvocalic coda consonant is voiced and the following consonant is voiceless. In practice, this means that a voiceless obstruent is geminated after a coda sonorant. I will refer to this special type of gemination as *Fortition*.

(24) Fortition:

a. *dansa* [dans.sa] 'to dance', *vänta* [vent.ta] 'to wait', *hjälpa* [jelp.pa] 'to help', *önska* [öns.ska] 'to wish', *minsta* [mins.sta] 'the least'

b. *dans* [danss] 'dance', *vänt* [ventt] 'turned', *valp* [valpp] 'puppy', *stark* [starkk] 'strong', *flöjt* [flöjtt] 'flute', *paus* [pauss] 'pause', *salt* [saltt] 'salt', *trumf* [trʉmff] 'trump', *skämt* [ʃemtt] 'joke', *Ulf* [ʉlff] (name), (W. Nyland) [skarffit] 'sharply' (Selenius 1972: 90)

The phonological nature of gemination Gemination applies only in stressed syllables, including those with secondary stress. Particularly interesting in this respect are the dialects of western Nyland, which have adjacent stressed syllables in a class of native and borrowed words (most with "grave" accent in West Swedish). Each of the stressed syllables undergoes Coda Gemination or Fortition, as the case may be (Selenius 1972: 94):

(25) [gamm.lasst] 'oldest' (from [ga.mal] 'old'), [tons.sill] 'tonsil', [kont.takkt] 'contact', [portt.följd] 'wallet', [bann.diit] 'bandit', [porss.liin] 'porcelain'

The asymmetry between stressed and unstressed syllables must be due either directly to Stress-to-Weight, which requires stressed syllables to be heavy, or indirectly to the inhibitory effect of Weight-to-Stress on lengthening of unstressed syllables (for these constraints see Prince and Smolensky 1993; Anttila 1997*a*; Kager 1999). I will pursue the latter approach, and posit general constraints corresponding to Fortition and Coda Gemination, dominated by syllabic well-formedness constraints, and by Weight-to-Stress, which requires heavy syllables to be stressed. High-ranking Dep-Stress prevents satisfaction of Weight-to-Stress by stressing, so gemination is blocked instead.[14]

If Stress-to-Weight or Weight-to-Stress are what restricts Fortition and Coda Gemination to stressed syllables, then these processes must increase syllable weight. Therefore they must add a mora to the syllable, which means that the lengthened consonant has the status of a true geminate. This is the first argument.

A convergent argument is based on the generalization that Coda Gemination does not apply after long vowels:[15]

(26) a. *vikta* /viik-t-a/ [viik.ta] 'folded' (pl.), not *[viikk.ta]

b. *bord* [buur(d)] 'table', not *[buurr(d)], *vald* [vaal(d)] 'elected' not *[vaall(d)] (parentheses indicate the weightlessness of final -C)

For, if Coda Gemination adds a mora, we can understand why it doesn't apply in (26), where the output of lengthening would be a four-mora syllable (taking final weightlessness into account in (26b)), a highly marked type. If, on the other hand, we were to suppose that Coda Gemination does not add a mora (but merely a non-moraic rhyme slot), we could not explain its failure to apply in (26), for syllables with four rhyme slots are quite common in Swedish, e.g. /viik-t-s/ [viikts] 'folded' (supine).

The force of the argument is somewhat weakened by the fact that Fortition does apply even in medial CVVC and final CVVCC syllables:[16]

[14] That Dep-Stress is undominated at the word level is independently motivated by the "stress-neutral" character of the word phonology.

[15] Cases like (Helsinki Swedish) *vakna* [va:k:na] 'to wake up' are not exceptions to this generalization. They arise not by Coda Gemination but by postvocalic Fortition (see below), based on the syllabification [vaak.kna], which the sonority profile allows.

[16] In Swedish, vowels are obligatorily lengthened before /rn/ and /rd/, as in *varna* [va:ɳa], *mord* [muuɖ] 'murder', and long vowels also occur in some words before /rt/, as in (27); this lengthening takes

(27) a. *karta* [kaart.ta] 'map'
b. *fart* [faart(t)] 'speed', *valt* [vaalt(t)] 'elected' (neuter)

However, the generalization about Coda Gemination remains striking. I tentatively conclude that Fortition and Coda Gemination are driven by distinct constraints, ranked in that order, with an intervening prosodic constraint which bars VVCC rhymes.

Both gemination processes are normally confined to the word domain.[17] This indicates (on our theoretical assumptions) that they are word-level processes, therefore phonological rather than phonetic. On the assumption that the phonological representation of quantity is moraic, this constitutes another argument for the proposed interpretation.

The upshot is that lengthening in Fenno-Swedish is genuine gemination, which adds a mora to a stressed syllable at the word level. Thus, in the lexical phonology, (23) and (24) are syllabified as, e.g., /viss.na/, /vikk(t)/, /dans.sa/, /dans(s)/, all with superheavy stressed syllables (parentheses mark weightless final consonants).

The scope of Fortition The Fenno-Swedish dialect of Borgå (Porvoo) does not have Fortition at all. Instead, it just lengthens the postvocalic coda consonant, even in words like (28) (contrast (24)).[18]

(28) [skvall.pas] 'to be splashed', [ʧörr.kan] 'the church', [grʉnnt] 'shallow' (Borgå; Harling-Kranck 1998: 26–8)

A number of dialects have Fortition not only postconsonantally but also after vowels.[19] The Swedish of Helsinki and Turku, the dialects of Åland, and the island of Nagu in the southwest, are of this type. The following examples are from Föglö (Åland; Harling-Kranck 1998: 84–6):[20]

effect even in Fenno-Swedish dialects, where these clusters do not fuse into a single retroflex consonant. Retroflex consonants, although phonetically single consonants, count as two consonants for purposes of syllable weight (as well as for other phonological constraints), in accord with their underlying status as clusters, e.g. *konsert* [konsæ:r] or [konsæʈ] (respectively with /-r/ and /-rt/); [*konsæ:ʈ] or [*konsær] are impossible. Dialectally, the lengthening applies before some other combinations of a sonorant plus a voiced consonant, e.g. eastern Nyland *saand* 'sand', *haald* 'hold'.

[17] However, Fortition occasionally occurs across compound boundaries and even across external word boundaries, e.g., *den konsekvensen* [dèŋǩ.kon.se.kvéns.sen] 'that consequence' (Itkonen 1965), though this is rather exceptional (Reuter 1982: 101).

[18] The articulation of voiceless stops is noticeably lenis in these dialects, but no more so than in some others which do show the more common lengthening pattern of (23) and (24).

[19] Diphthongs seem to pattern with long vowels, e.g. Snappertuna (western Nyland) [poi.ki] 'boy'. In dialects with post-long vowel gemination, the voiceless stop would of course be geminated, e.g. Helsinki/Turku *pojke* [poik.ke].

[20] In the text from the island of Kökar in eastern Åland (Harling-Kranck 1998: 78–81), postvocalic Fortition is variable.

(29) Postvocalic Fortition:

a. /eeta/ [eet.ta] ‘eat’, /smaaka/ [smaak.ka] ‘taste’, /baaka/ [baak.ka] ‘bake’, /flaata/ [flaat.ta] ‘flat surface’

b. [maatt][21] ‘food’, [gröött] ‘porridge’

In these dialects, the medial consonants of words like *mata* [maat.ta] ‘to feed’, *kåkar* [kook.kar] ‘hovels’—phonemically singletons—are phonetically about as long as the underlying geminates of words like *matta* [mat.ta] ‘carpet’, *kockar* [kok.kar] ‘cooks’.[22]

Most Fenno-Swedish dialects don’t have postvocalic Fortition. In them, a word like *rita* ‘to draw’, phonemically /riita/, is pronounced [riita], nearly like Finnish *riita*. The short medial consonant in such words is a salient shibboleth of rural Fenno-Swedish.

The strict parallelism of final and non-final syllables with respect to Fortition across dialects constitutes more evidence for -C weightlessness. The following implications hold:

(30) a. Postconsonantal Fortition: [skvalp.pas] ⇔ [valpp], [skvall.pas] ⇔ [vallp]

b. Postvocalic Fortition: [maat.ta] ⇔ [maatt], [maa.ta] ⇔ [maat]

If -C is weightless, the processes can be unified. In our analysis, /skval.pas/ → /skvalp.pas/ is parallel to /val(p)/ → /valp(p)/, and /maa.ta/ → /maat.ta/ is parallel to /maa(t)/ → /maat(t)/.

6.1.4 *The syllabic typology of Fenno-Swedish dialects*

Six weight systems The syllable weight properties just reviewed—light stressed syllables, distinctive superheavy syllables, and redundant superheaviness due to Coda Gemination and Fortition in its two varieties—do not combine freely. In fact, just six basic quantitative systems are attested in Fenno-Swedish. These are tabulated in (31).[23]

[21] The gemination of word-final consonants is heard clearly when a vowel follows in close contact in the next word. Examples from the dialect recordings are [maatt ifroon] ‘food from’, [gröött o. . .] ‘porridge and. . . ’, [mjölkk o smör] ‘milk and butter’ (Harling-Kranck 1998: 85). Contrast [tibaak o] ‘back too’ (109), [maat o kaffe] ‘food and coffee’ (110), from a dialect without postvocalic Fortition (South Ostrobothnia).

[22] Intervocalic lengthening also occurs in Sweden (Elert 1965: 145, 186). There it is not quite as marked as in Helsinki, and I do not take a position on whether it should be analyzed as gemination, as in Fenno-Swedish. However, the lengthening is quite marked, and more than outweighs the lengthening of the vowel before voiced consonants: e.g., the overall duration of *rita* /riita/ ‘to draw’ is longer than the overall duration of *rida* /riida/ ‘to ride’ (Elert 1965: 162).

[23] The words in the table are meant to represent only quantitative *types*. Their actual vowel and consonant qualities may differ from dialect to dialect in ways that are irrelevant to the present discussion.

(31) Fenno-Swedish syllable types:

	General	S.Ostrob.	Borgå	S.W.	Helsinki	Brändö	
1. *baka*	[baka]	[baka]	[baka]	[baka]	[baakka]	[baakka]	'to bake'
2. *mina*	[mina]	[mina]	[mina]	[mina]	[mina]	[minna]	'my' (pl.)
3. *gått*	[goott]	[goott]	[goott]	[gott]	[gott]	[gott]	'gone'
4. *vända*	[vennda]	[vennda]	[vennda]	[vennda]	[vennda]	[vɛnnda]	'to turn'
5. *vänta*	[ventta]	[ventta]	[vennta]	[ventta]	[ventta]	[vɛntta]	'to wait'
6. *ropa*	[ruupa]	[ruupa]	[ruupa]	[ruuppa]	[ruuppa]	[ruuppa]	'to call'
7. *råg*	[rog]	[roog]	[rog]	[roog]	[roog]	[roog]	'rye'

Row 1 shows whether light stressed syllables occur in lexical words, and row 2 shows whether they occur in function words. The next four rows show, respectively, the distribution of lexically distinctive superheavy syllables (long vowel plus geminate consonant), regular Coda Gemination (common to all dialects), postconsonantal Fortition, postvocalic Fortition, and lexical CVC words (recall that lexical CV words are excluded everywhere).

The first column, labeled "General," represents the most common pattern, scattered throughout the Fenno-Swedish area from Nyland (Uusimaa) in the South, through part of the Southwest, and into central and northern Ostrobothnia in the North. The other dialects are confined to particular localities. South Ostrobothnia (column 2) and Borgå in Nyland (column 3) share the full contrast between light, heavy, and superheavy syllables. The remaining dialects lack contrastive superheavy syllables (columns 4–6). In addition, Helsinki/Turku (column 5) has light stressed syllables and CVC words only under limited conditions (as discussed below), and Brändö (on Åland) lacks them completely. Abstracting away from particulars, then, the typology can be schematized as follows:

(32)

	General	S. Ob.	Borgå	S.W.	Helsinki	Brändö
1. light stressed syllables	yes	yes	yes	yes	(yes)	no
2. lexical superheaviness	yes	yes	yes	no	no	no
3. postvocalic Fortition	no	no	no	no	yes	yes
4. postconsonantal Fortition	yes	no	no	yes	yes	yes
5. CVC words	yes	no	yes	no	(yes)	no

Three generalizations emerge from (31) and (32).

- Postvocalic Fortition implies postconsonantal Fortition.
- Postvocalic Fortition is incompatible with contrastive superheaviness.[24]
- Postvocalic Fortition is incompatible with lexical light stressed syllables.

An attempt to explain the distribution of syllable types and the above implicational generalizations follows. It is based on a synchronic phonological analysis in terms of the Stratal OT model. By way of preface, a few remarks on the origin of Fenno-Swedish gemination are in order.

6.1.5 *The origins of Fenno-Swedish syllable structure*

Itkonen (1965) and Reuter (1982) theorize that the characteristic quantitative properties of Fenno-Swedish are the result of accommodation to one of the two quantitative models available in Finnish words. Consider a word like *rita* 'to draw', phonologically /riita/, in Sweden pronounced [riit:a], with a lengthened stop. In the Fenno-Swedish dialects without intervocalic gemination, it is pronounced like Finnish *riita* 'discord' (CVVCV). In the educated urban Swedish of Helsinki and Turku, it is pronounced just about like Finnish *Riitta* (CVVCCV). According to Itkonen and Reuter, this dialect split within Fenno-Swedish arose because native speakers of Finnish acquiring Swedish could identify the phonemically short, but phonetically lengthened intervocalic voiceless obstruents of Swedish either with the short consonants of Finnish (giving rise to the majority of dialects) or with the long consonants of Finnish (Helsinki, Turku, SW islands).[25]

Still, we have to ask *why* the dialects have split this way. Why did they not all choose gemination, which better approximates the West Swedish pronunciation? The reason why most dialects did not adopt postvocalic

[24] Harling-Kranck 1998: 155 cites the form *sjöött* from Finström in Åland, a dialect with post-long vowel gemination, which would be the sole exception to this generalization in the entire collection of dialect material. However, this citation seems to be an error. In the actual text, transcribed in two versions, as well as the accompanying recording, this word clearly has a short vowel.

[25] The borrowing of Swedish words into Finnish usually reflects both intervocalic gemination and cluster gemination. For example, the Swedish name *Brita* is rendered as *Riitta* in Finnish, as would be expected if it were taken from a dialect with post-long vowel gemination. The Swedish word *simpel* is rendered as *simppeli* in Finnish, as would be expected if it were taken from a dialect with cluster gemination. (For some reason, gemination of fricatives in borrowings is not so regular; Reuter 1982: 154 ff.) The pattern was presumably established on the basis of the Fenno-Swedish prestige dialect, which has both these gemination processes. Since then, gemination has simply become a conventional way of rendering foreign voiceless stops in Finnish, even when they are not actually geminated in the source language. For example, in *pankkiiri* 'banker', Finnish has a geminate even though the Swedish source word *bankir* [baŋkíir] has a singleton (because the preceding vowel is unstressed), and the Finnish spoken-language rendition of 'Clinton', *Klinttoni*, has a geminate even though the English source has a singleton.

Fortition—in terms of the substratum theory, why their speakers interpreted the Swedish lengthened postvocalic voiceless obstruents as singletons—may be that (except for South Ostrobothnia) these dialects have lexically distinctive superheavy syllables. The generalization is that Fortition was avoided wherever it would have merged a contrast between heavy and superheavy syllables. This would reflect a functional principle of *contrast preservation* (Flemming 1995, 2001; Padgett 2003). If we suppose that South Ostrobothnia shortened its superheavy syllables after the gemination system was established, we would even have the stronger generalization that Fortition was introduced wherever possible to enhance heavy syllables provided the distinction between heavy and superheavy syllables was not suppressed.

Two further facts lend support to this scenario. It explains an otherwise puzzling asymmetry between the two Fortition environments. Few dialects have postvocalic fortition, whereas all dialects except for Borgå have consonantal Fortition. From the contrast preservation perspective the explanation is obvious. Postconsonantal geminates are never contrastive in Swedish, so contrast preservation is irrelevant to them, and speakers were free to choose the phonetically closest rendition as geminates.

Perhaps the most striking evidence comes from monosyllabic words. In the dialects that maintain the distinction between CVC and CVCC words, both Coda Gemination and Fortition are obviously inapplicable to monosyllabic words—otherwise they would surface as CVCC. Restricting Gemination and Fortition to polysyllabic words would however be unnatural and stipulative. In any case, the reason the CVC : CVCC contrast is retained is because CVC words escape vowel lengthening due to the low ranking of C-Extrametricality, as shown in (7). The generalization that Gemination and Fortition do not neutralize any contrasts extends to these cases as well, however.

The non-neutralizing property of the gemination processes is also relevant to the synchronic analysis, to which I now turn. I will argue that it should be factored out into a general anti-neutralization constraint.

6.1.6 *The gemination system*

The constraints Let us suppose that gemination is effected by two constraints.

(33) a. Fortition: A voiceless consonant is geminated.

b. CodaGemination: A postvocalic coda consonant is geminated.

FORTITION and CODAGEMINATION are probably to be decomposed into more elementary constraints, but I will not pursue this refinement further here. The contextual restrictions on them emerge from higher-ranked constraints. For example, syllable structure constraints prohibit Fortition in onsets. The mutual ranking of FORTITION and CODAGEMINATION and their ranking with respect to other constraints determine the dialectal variation with respect to gemination. These include the prosodic maximality constraint $^*\mu\mu\mu$ (see (22b)), and, more interestingly, a synchronic NONEUTRALIZATION constraint, the counterpart to the diachronic explanation for the dialectal distribution and contextual restrictions on gemination explored in the preceding section.

In standard OT phonology the expectation is that the system of lexical contrasts should emerge from the constraint system. A constraint which prohibits neutralization turns this backwards. The argument for such a constraint is that it allows several generalizations to be captured which are otherwise lost. First, it explains why postvocalic Fortition does not apply in any dialect where /CVVC/ (and /CVVC(C)/ in monosyllables) is distinctive: for in just those dialects it would wipe out a lexical contrast. Notice that in this case the direction of explanation cannot be reversed. That is, we cannot attribute the absence of distinctive /CVVC/ in West Swedish and Helsinki to postvocalic Fortition, for several reasons. First, the neutralization applies equally before *voiced* consonants, where Fortition is inapplicable. Secondly, the neutralization is in fact not effected by Fortition, but by shortening of /CVVC/ to /CVC/, e.g. [ruudde] > [rudde] 'rowed' (past), [ruutt] > [rutt] (pp.). Therefore it is the existence of distinctive /CVVC/ (due to the stem-level ranking MAX$\mu \gg {}^*\mu\mu\mu$) that constrains Fortition, not the other way round.

A similar argument is based on dialects that distinguish /CVC/ words from /CVCC/ words. The explanation cannot involve merely restricting Coda Gemination and Fortition to polysyllables, for, as shown in (7), the primary cause of the retention of the CVC : CVCC contrast is the low ranking of C-EXTRAMETRICALITY, which allows CVC words to escape vowel lengthening. Conversely, the generalization that Gemination and Fortition do not neutralize any contrasts extends to these cases as well.

This justifies a constraint which prevents gemination from erasing weight contrasts. The most general formulation would be NONEUTRALIZATION:

(34) NONEUTRALIZATION: An output must not have a more faithful input correspondent.

An output $\mathcal{A}$ corresponding to input A violates NONEUTRALIZATION if there is an input B such that B $\Leftrightarrow$ $\mathcal{A}$ incurs fewer faithfulness violations than

A ⇔ 𝒜. The effect of NoNeutralization in General Fenno-Swedish (type (8)) is summarized in (35).

(35) a. /riita/ ↛ *[riitta] (Fortition is blocked because [riitta] has the more faithful input correspondent /riitta/)

b. /led/ ↛ *[ledd] (Coda Gemination is blocked because [ledd] has the more faithful input correspondent /ledd/)

c. /las/ ↛ *[lass] (Coda Gemination and Fortition are blocked because [lass] has the more faithful input correspondent /lass/)

In each case, the output candidates incur a violation of NoNeutralization at the word level because they have input correspondents which have fewer faithfulness violations. This will be true for all dialects which admit distinctive superheavy syllables. Similarly, in the dialects with a /CVC/ : /CVCC/ opposition, Coda Gemination and Fortition of /CVC/ to [CVCC] incur a violation of NoNeutralization, because the output [CVCC] has a more faithful input correspondent /CVCC/.

Let us suppose that (34) is formally like any other constraint in that it can be ranked with respect to the other constraints. This means that at any given level the markedness constraints will divide into those that can effect neutralization and those that cannot, with the two sets separated by (34).

Stratal OT The alternative inputs to which NoNeutralization refers do not have to be actual lexical items, just *possible* inputs. This presupposes some way of characterizing possible inputs independently of the constraints that map inputs to outputs. In fully parallel OT, such a characterization is not available because, under the Richness of the Base assumption, *any* input form is admissible. The form, if underlying representations, emerges from the constraint system itself via Lexicon Optimization. Thus, constraints such as NoNeutralization, which refer to possible inputs, are not available in parallel OT.

However, I have argued on independent grounds that parallel OT should be rejected (Kiparsky 2000, 2002, forthcoming). Instead, I propose to adopt Lexical Phonology's distinction between lexical and postlexical phonology, where the lexical phonology itself comprises a stem phonology ("level 1") and a word phonology ("level 2"). (It goes without saying that this organization is not specific to Swedish but common to all languages.) Contrary to traditional Lexical Phonology, however, I view each of these phonological subsystems as a parallel OT constraint system. These constraint systems may differ in ranking. All seriality lies in the interface between the levels. Within the lexical phonology, the output of the stem level is the input to the word level:

(36)

Stem-level constraints
↓
Word-level constraints

The output of the word level is in turn the input to the postlexical constraint system.

I'll call this marriage of OT and Lexical Phonology Stratal OT (a term suggested by John McCarthy). The major arguments for Stratal OT, that it provides a unified, restrictive, and simple treatment of phonological opacity and cyclicity, have been presented elsewhere. Its significance for the present study of Swedish word phonology is that it allows us to distinguish between the quantitative restrictions on stems and those on words. Because the levels interface serially, words derived from stems inherit the latter's quantitative properties insofar as the word phonology permits. The two specific corollaries that we will be exploiting are the possibility of characterizing the class of possible inputs to the word phonology, and distinguishing in a principled way between lexical words and function words.

To summarize: from the OT perspective, a contrast is absent when the faithfulness constraints that would maintain it are dominated by the markedness constraints that suppress it. Under parallelism, contrast is definable only on output representations. In Stratal OT, contrast is definable on the output of each phonological level. A contrast which exists at one level might be neutralized by a markedness constraint at another. At the stem level, Richness of the Base and Lexicon Optimization figure exactly as in parallel OT (in this respect no different from the traditional approach of Lexical Phonology). The inputs to the word level are just the outputs of the stem level, with word-level morphology applied. Constraints such as NoNeutralization, which make reference to what is a possible input, are therefore definable. For example, /CVC/ is a possible input to the word phonology in a given dialect of Swedish just in case it is a possible output of the stem phonology in that dialect. This provides a straightforward way to define neutralization and contrast preservation.

The general Fenno-Swedish pattern of gemination is obtained by the word-level ranking shown in (37):

(37) NoNeutralization $\gg$ Fortition $\gg$ CodaGem $\gg$ $^*\mu\mu\mu$

As can be seen in (64), the ranking Fortition $\gg$ CodaGemination is crucial in cases like *vänta* 'wait', which is pronounced [vent.ta], not *[venn.ta].

FORTITION and CODAGEMINATION converge in words like *räkna* 'to count', *atlas* 'atlas', pronounced [rekk.na], [att.las]. When both consonants of the cluster are voiced, as in *vända* 'turn', *semla* 'bun', *Selma* (proper name), FORTITION is not at stake, so (33b) CODAGEMINATION requires lengthening the postvocalic consonant (rather than the onset): [venn.da], [semm.la], [sell.ma], not [*ven.dda], *[seml.la], *[selm.ma]). When all consonants of a cluster are voiceless, then syllable structure allows only one of them to be geminated; by CODAGEMINATION this is the postvocalic one, so *hetsar* 'incites', *hästar* 'horses', are pronounced [hett.sar], [hess.tar] (not *[hets.sar], *[hest.tar]). In viewing the tableau, keep in mind that this being the word phonology, the inputs are the stem-level outputs. Observe the role of NONEUTRALIZATION in items 6, 7, and 8.

(38)

General F.-Sw. (W. L.)	NONEUTR	FORTITION	CODAGEM	*$\mu\mu\mu$
Input: /rista/ 'to carve'				
1a. rís.ta		**	*	
1b. ☞ ríss.ta		*		*
1c. ríst.ta		*	*	*
Input: /velja/ 'to choose'				
2a. vél.ja			*	
2b. ☞ véll.ja				*
2c. vélj.ja			*	*
Input: /ven.da/ 'to turn'				
3a. vén.da			*	
3b. ☞ vénn.da				*
3c. vénd.da			*	*
Input: /venta/ 'to wait'				
4a. vén.ta		*	*	
4b. vénn.ta		*		*
4c. ☞ vént.ta			*	*
Input: /riida/ 'to ride'				
5a. ☞ ríi.da				
5b. ríii.da				*
5c. ríid.da	*			*
Input: /riita/ 'to draw'				
6a. ☞ ríi.ta		*		
6b. ríii.ta		*		*
6c. ríit.ta	*			*

Input: /stöött/ 'hit' (pp.)					
7a. ☞	stöött				*
7b.	stött	*			
7c.	stööt	*	*	*	
Input: /las/ 'read'					
8a. ☞	las		*	*	
8b.	las(s)	*			

The Borgå ranking differs only in that CodaGemination is undominated (its ranking with respect to NoNeutralization is immaterial), so that it swamps out any visible effect of Fortition:

(39) NoNeutralization, CodaGem $\gg$ Fortition $\gg$ Max-μ $\gg$ $^*\mu\mu\mu$

(40)

Borgå (W. L.)		NoNeutr	CodaGem	Fortition	$^*\mu\mu\mu$
Input: /rista/ 'to carve'					
1a.	rís.ta		*	**	
1b. ☞	ríss.ta			*	*
1c.	ríst.ta		*	*	*
Input: /velja/ 'to choose'					
2a.	vél.ja		*		
2b. ☞	véll.ja				*
2c.	vélj.ja		*		*
Input: /ven.da/ 'to turn'					
3a.	vén.da		*		
3b. ☞	vénn.da				*
3c.	vénd.da		*		*
Input: /venta/ 'to wait'					
4a.	vén.ta		*	*	
4b. ☞	vénn.ta				*
4c.	vént.ta		*		*
Input: /riida/ 'to ride'					
5a. ☞	ríi.da				
5b.	ríii.da				*
5c.	ríid.da	*			*
Input: /riita/ 'to draw'					
6a. ☞	ríi.ta			*	
6b.	ríii.ta			*	*
6c.	ríit.ta	*			*

Input: /stöött/ 'hit' (pp.)					
7a. ☞	stöött				*
7b.	stött	*			
7c.	stööt	*	*	*	
Input: /las/ 'read'					
8a. ☞	las		*	*	
8b.	las(s)	*			

So far our analysis does not incorporate the grammatical constraints on the distribution of light stressed syllables that we noted for Helsinki, specifically the fact that they occur just in function words and in a few other small classes of lexical items. The following section supplies the missing pieces necessary for understanding this grammatical conditioning. It amounts to an independent argument for Stratal OT.

6.2 Stems and words

6.2.1 *Light stressed syllables in Helsinki*

Helsinki (and Turku) Swedish has light stressed syllables in the following classes of words:

(41) a. in function words before voiced consonants,
  b. in words where the open syllable results from epenthesis,
  c. in a class of suppletive verb allomorphs,
  d. in truncated words.

Otherwise they occur only in a small number of polysyllabic words (mostly Finnish loans). The environments in (41) seem like a motley assortment, but we shall see that they have something interesting in common that explains why they go together.[26]

[26] Even more limited CV́ appears in some of the dialects of Åland. For example, *bara* 'only', *seta* 'to put', *rikit* 'really', *såna* 'such' in Kökar (Harling-Kranck 1998: 78 ff.), and *kuna* 'to be able', *bara* 'only', *någe* 'some' in Saltvik (*ibid.* 88 ff.). Except for *seta*, and perhaps *kuna*, these words have (or can have) short vowels in Helsinki/Turku also. However, most words which have light stressed syllables in the latter dialects seem to have geminate consonants in Åland, e.g. *minna* 'my' (pl., Helsinki *mina*, West Sw. *mi:na*), *meddan* 'while' (Helsinki *medan*, West Sw. *me:dan*). Occasionally even Sweden goes with Åland in having geminates in place of the Fenno-Swedish light stressed syllables: Helsinki *honom*, West Sw. *honnom* 'him'; Helsinki *i moron*, West Sw. *i morron* 'tomorrow'; Helsinki *hade* ∼ *hadde* 'had', West Sw. *hadde* 'had'.

In what follows, words cited in italics represent standard Helsinki Swedish in regular Swedish spelling, which marks distinctive consonant length by gemination. I add colons to mark vowel length, and (where necessary) primary and secondary accents, and periods to mark syllable boundaries. The reader can easily recover the actual Swedish orthography by just erasing these marks. The actual pronunciation can be recovered as far as syllable weight is concerned by applying Fortition and Coda Gemination under the conditions stated above, and for the vowels, by the correspondences *å* = [o], *o* = [u], *u* = [ʉ], *ä* = [e].

A list of function words with short stressed syllables in Helsinki Swedish is given in (42). Observe that consonant immediately following the short CV́ syllable is always voiced.

(42) a. *Pronouns, determiners: ja* 'I', *du* 'you', *vi* 'we', *ni* 'you', *honom* 'him', *de* 'it', *va* 'what', *mina* 'mine (pl.)', *dina* 'your (sg.)', *våra* 'our', *era* 'your (pl.)', *deras* 'their', (but *henne* 'her' (acc.)), *någo* 'something', *hudan*, *váför en* 'what kind of', *sådan*, pl. *såna* 'such', *sånahána* 'this kind of', *sånadána* 'that kind of', *hela (da:gen)* 'all (day)'

b. *Auxiliaries: ha* 'have', *hade* ∼ *hadde* 'had', *ä* 'is', *va* 'was', *ska* 'shall', *vara* 'be', *vari(t)* 'been', *blivi(t)* 'become', *sku* 'should', *sku boda* 'should have', *sku vila* 'would like to', (but *må:* 'may')

c. *Prepositions, particles, verb prefixes: före* 'before', *genom* 'through', *över*, *inom* 'within' (vs. *mellan* 'between', *u:tan* 'without')

d. *Conjunctions: å* 'and', *både* 'both [...and]' (but *bå:da* 'both', determiner), *medan* 'while', *bara* 'if only' (vs. *innan* 'before')

e. *Small adverbs:*[27] *så* 'so', *då* 'then', *nu* (1) 'now', (2) affirmative (= West Sw. *no:g*), and their derivatives: *númè:ra*, *núförtì:den* 'nowadays', *dåförtì:den* 'in those days', *me* 'too', *ändå* 'still' (can be end-stressed), *bara* 'only', *redan, ren* 'already', *igenom* 'through', *óvan* 'above, over'

f. *Complementizer: å* 'to' (infinitive purpose clauses)

g. *Interjections: jahá* 'I see', *ahá* 'aha', *nå* 'nu', *tja* 'well'

The systematic character of the restriction to function words is underscored by the fact that, when function words are promoted to lexical words, any stressed light syllables in them are automatically lengthened, in conformity with the regular quantitative constraints on stems.

[27] This class was identified for Finnish in Hanson (1992) and Hanson and Kiparsky (1996: 320) as adverbs which "constitute entire phrases and so permit no modification or complementation." A general theory of such "non-projecting categories" is presented in Toivonen (2001).

(43) a. *nu* 'now' vs. *nu:-et* 'the present time'
b. *mina* 'my (pl.)' vs. *de mi:na* 'my relatives, my loved ones'
c. *hela (da:gen)* 'all (day)' vs. adj. *(en) he:l (da:g)* '(a) whole (day)', and *he:l, he:l-a* 'entire, undamaged'
d. *ja* 'I' vs. *ja:g-et* 'the ego'
e. *ha* 'have' (auxiliary) vs. *att ha:* 'to have'

6.2.2 *Explaining the distribution*

What is the basis for the phonological distinction between lexical words and function words? Stratal OT interprets Lexical Phonology's "level 1" and "level 2" as stems and words, and takes their respective phonologies to be governed by distinct constraint systems. Because the levels interface serially—that is, the output of the stem phonology is the input to the word phonology—words derived from stems inherit the latter's properties insofar as the word phonology permits.

Lexical words (nouns, verbs, adjectives, adverbs of the projecting type) enter the derivation as stems, while function words do not. Therefore, stems must conform to an additional set of phonological constraints, namely those which constitute the stem phonology. But both lexical words and function words are subject to the word phonology, and both participate in the postlexical phonological derivation.[28] Moreover, the templatic truncation morphology is also demonstrably a word-level process.

In Helsinki Swedish, stressed syllables are strictly bimoraic in the stem phonology. In the word phonology, one-mora syllables arise through function words, truncation, and epenthesis, and three-mora syllables arise through gemination.

(44)

Stem phonology:
Stressed syllables have exactly 2 moras
Vowel lengthening and shortening

↓

Word phonology:
Stressed syllables have at most 3 moras
Gemination

[28] In Kiparsky (forthcoming) I provide independent evidence for this claim from a number of languages. For example, function words in English are not subject to lexical stress, to Vowel Shift, or to Philadelphia æ-"tensing." Cross-linguistically, it is well known that roots and function words are not necessarily subject to the same prosodic minimality constraints as words are.

Formally, in the stem phonology the prosodic constraints (6c) Stress-to-Weight which requires stressed syllables to have at least two moras, and (22b) $^*\mu\mu\mu$, which prohibits syllables of more than two moras, both outrank the faithfulness constraints that prevent vowel lengthening and shortening (Max-V, Dep-V). In the word phonology, however, the prosodic constraints are outranked by Fortition and CodaGemination, as well as by faithfulness constraints. Thus, superheavy and light stressed syllables are prohibited in stems, but not in words.

Function words According to our proposed analysis, surface CV́ syllables occur just in words which for some reason escape lengthening at the stem level. These turn out to be just the four types of words with light stressed syllables in (41). The simplest case is that of function words. By hypothesis, function words are not stems, therefore not subject to stem phonology. Of course, they are words, and as such subject to word phonology. But lengthening is enforced only in the stem phonology, not at the word level. Therefore, function words retain underlying short syllables even under stress.

Once again, CVC monosyllables pattern like CV́ in polysyllabic words: they occur only in function words, and only where -C is voiced. The contrast between final single and geminate stops tends to be neutralized in citation forms, but it is audible within a phonological phrase, particularly when a vowel follows:

(45) a. /ann/, /hann/: *om Ann inte hann ä:ta* [om ann int hann eetta] 'if Ann didn't have time to eat'

b. /han/, /kan/: *om han inte kan ä:ta* [om han int kan eetta] 'if he can't eat'

These data suggest that Coda Gemination in these dialects applies only in the postlexical phonology.

Epenthesis Case (41b) comprises words which are underlying monosyllables of the form /CVCL/, pronounced as monosyllabic before vocalic endings and as disyllabic elsewhere in virtue of epenthesis of *-e-* to break up the final cluster. These words retain underlying light syllables before voiced consonants, resulting in the three-way surface contrast between /CV́-/, /CV́C-/ and /CV́V-/ seen in (46). Before voiceless consonants, we just get the usual two-way distinction between /CV́C-/ and /CV́V-/.

(46) a. *hy.vel* 'plane' (tool), (pl. *hyv:.lar*, *hyv:.la* 'to plane'),[29] *stö.vel* 'boot' (pl. *stöv:.lar*), *ö.verst* 'uppermost' (*öv:.re* 'upper'), *ö.ver.ste* 'colonel'

b. *kä:.gel* 'bowling pin' (pl. *kä:.glar*), *få:.gel* 'bird', (pl. *få:.glar*), *i:.gel* 'leech' (pl. *i:.glar*), *sni:.gel* 'snail' (pl. *sni:.glar*), *na:.vel* 'navel' (pl. *na:v.lar*), *spektá:kel* [spek.tá:k.kel] 'spectacle' (pl. *spek.tá:k.kler*)

c. *dub.bel* 'double' (pl. *dub.bla*), *nyckel* [nyk.kel] 'key' (pl. *nyck-lar*) [nyk.klar], *cykel* [syk.kel] 'bicycle' (pl. *cyklar* [syk.klar]), *smug.gel(go:ds)* 'smuggling, contraband' (*smug.gla*), *spug.gel* 'barf' (*spug.gla*)

In a stem such as /hyvl/, the conditions for lengthening are not met, and the vowel stays short (as in other words ending in -CC, e.g. *kalv* 'calf'). Epenthesis takes place just at the word level. The evidence is that it is bled by vowel-initial suffixes (e.g. inflection), as the examples in (46) show. But lengthening does not apply to words. Therefore, an underlying short vowel is retained even when it comes to stand in a final open syllable by epenthesis.

Root inflections Case (41c) is represented by a small class of lexical words with light stressed syllables which are inflected from bound roots. The periphrastic perfect is based on the so-called supine form, which is normally built on the verb stem, in which case it conforms to the lexical length constraints (e.g. inf. *veta* [ve:t.ta] 'know', supine *vetat* [ve:t.tat]). Some verbs, however, can form their supines from a bound root form. For example, the verb 'to strike' has a supine from a bound root form *slaj-*, which is not used in any other form of the verb. Just the root-based inflections stay short; contrast the other forms in (47), which lengthen regularly:

(47) *dragit* [dra.ji] 'pulled' (pres. *dra:*, past *dro:g*), *slagit* [sla.ji] 'hit' (*slå:, slo:g*), *tagit* [ta.ji] 'taken' (*ta:, to:g*), *givit* [ji.vi] 'given' (*ge:, ga:v*), *blivit* [bli.vi] 'become' (*bli:, ble:v*)

The supines of the first three verbs can also be formed from regular verb stems, in which case they have the expected long vowel, e.g. *dra:gi(t), sla:gi(t)* (stem *dra:g, sla:g*). Elsewhere, lengthening applies regularly to these verbs. For example, 'to strike' has the stems *slå:* (*slå:.en.de* 'striking', *sla:g* 'a strike', *slo:g* 'struck').

[29] A reminder: cited words in italics are in Swedish spelling, with periods added to mark syllable boundaries, and macrons to mark vowel length and tautosyllabic consonant length. The spelling of these words is *hyvlar, hyvla*.

Treating these suppletive forms as inflected roots immediately accounts for their short vowel. In particular, a root such as *draj-* is not subject to lengthening. The output of adding the supine suffix *-it* at the word level undergoes *word-level* phonology, where lengthening is not operative. It is a long-standing assumption of Lexical Phonology that bound roots are not "cyclic domains," i.e. that they are phonologically inert in themselves, and undergo phonology only in combination with affixes.[30]

Truncated words To appreciate the last class of cases, an additional generalization must be understood: that the two-mora minimum on stressed syllables is enforced only in *non-final feet*. In fact, a general process of prestress shortening and destressing (which also applies to some extent in West Swedish dialects) leads to alternations such as the following:

(48) a. *systé:m* 'system'
systemá:tisk 'systematic' [sys.te.máat.tisk]
systematí:k 'systematism'
systematisé:ra 'systematize'

b. *tjä:nare* 'servant', *tjä̀narínna* 'female servant'
gu:d 'god', *gudínna* 'goddess'
gre:ve 'count', *grevínna* 'countess'

In Fenno-Swedish the process is considerably more general:

(49) a. *tjùgusjú:* 'twenty-seven' (*tju:gu* 'twenty')

b. *várifrå̀:n* 'from where', *dä́rifrå̀:n* 'from there', (*va:r* 'where', *dä:r* 'there')

In long (mostly foreign) words, only a final (binary or unary) foot regularly requires its stressed syllable to be heavy. Syllables in non-final feet, whether bearing primary or secondary stress, remain short, regardless of the voicing of the following consonant.

(50) a. *kválitatì:v* 'qualitative', *pósitì:v* 'positive', *hýperkorrèkt* 'hypercorrect', *póliklì:nik* 'clinic', *sémikò:lon* 'semicolon', *nóminatì:v* 'nominative', *génetì:v* 'genitive', *élatì:v* 'elative', *íteratì:v* 'iterative', *féminì:n* 'feminine', *décilì:ter* 'deciliter', *géneratì:v* 'generative', *mínikjò:l* 'miniskirt', *sémikò:lon* 'semicolon', *Fö́lisö̀:n* (place name). Many of these can also have stress on the final foot, e.g.

[30] The reason is assumed to be that bound roots are not prosodified (and in fact do not need to meet prosodic minimality constraints), see in general Inkelas (1989).

[pà.pe.gój:.ja], including most words in *-iv*, e.g., *prìmití:v* ‘primitive’, *rèlatí:v* ‘relative’, etc.

b. *krèatú:r* ‘creature’, *tèologí:* ‘theology’, *tèoló:g* ‘theologian’, (*fìloló:g, pèdagó:g*, etc.), *pàradí:s* ‘paradise’, *pàragrá:f* ‘paragraph’, *tèlegrá:f* ‘telegraph’, *èpidemí:* ‘epidemic’, *àkademí:* ‘academy’, *sỳnagó:ga* ‘synagogue’

c. *kámera* ‘camera’, *dómino* ‘domino’, *dómina* ‘domina’, *númerus cláusus* ‘quota’, *mínimum* ‘minimum’, *dýnamo* ‘dynamo’, *ánanas* ‘pineapple’, *sýfilis* ‘syphilis’, *plátina* ‘platinum’, *stímulus* ‘stimulus’, *faksímile* ‘facsimile’, *Távaststjèrna, Ágaton, Kásimir, Sálomon, Júpiter* (personal names), *Ládoga, Árarat, Távastlànd* (place names)

In the last set of cases the CV́CV foot is non-final in virtue of being followed by another syllable. Thus, non-final feet do not become superheavy.

Under secondary stress, closed syllables are also lengthened by gemination of voiceless consonants, as in (51a,b,c); contrast (51d,e):

(51) a. *elak* [ée.làkk] ‘evil, nasty’, *elaka* [ée.làk.ka] (pl., def.)
b. *palsternacka* [páls.ter.nàk.ka] ‘parsnip’
c. *enstaka* [éen.stàak.ka] ‘sporadic’
d. *nutida* [nʉ́ʉ.tìi.da] ‘contemporary’
e. *idog* [íi.dùug] ‘diligent’, *idoga* [íi.dùu.ga] (pl., def.)

We are now ready for case (41c). When long words of the type just examined get truncated, their initial foot becomes word-final, *but the truncated form still retains its CV́ syllable*, this time irrespective of the voicing of the following consonant.

(52) *foto* ‘photograph’ (from *fòtografí:*), *Tele(-verket)* ‘the phone company’ (from *tèlefó:n)*, *kilo* ‘kilo(gram)’ (from *kìlográm*), *Hypo(banken)* ‘Mortgage Bank’ (from *hỳpoté:k*), *día(bild)* ‘slide, transparency’ (from *díapòsitì:v*), *Bío-Bìo* (a movie theater), (from *bìográ:f* ‘movie theater’) (but *bi:o* ‘movie theater’), *Póli* (from *(Pòly)tékniska Högskolan* ‘Polytechnical University’), *Majo* ‘the Majority’ (from *Màjorité:ten* ‘the Majority’, a grass-roots citizen’s organization, Reuter 1986)

In the stem-level representation, the base begins with a light syllable. Truncation is a word-level process, as shown by the fact that it applies to words with the postposed definite article to make an inherently definite truncatum, as in *Tele, Hypo, Poli, Majo* (see (52)). Lengthening is not applicable at the word level, therefore in particular not to truncated words. It follows that an underlying short vowel is retained even when truncation puts it into the word-final foot.

Exceptions and residual cases Akin to truncations are lexicalized fast speech forms reduced to CVCV form by simplification of medial clusters. They too retain the short vowel of the original (regardless of voicing).

(53) *rikit* 'really' (from *rikti(g)t*), *vika, viken* 'which' (from *vilka, vilka*), *moron* 'morning' (from *morgon)*

They are no longer outputs of a productive reduction process, but are simply lexicalized with an underlying short vowel.

There remains a small set of words with unexplained CV́.[31]

(54) *göra* [jö.ra] 'do', *käring* 'old woman', *senap* 'mustard', *tobak* 'tobacco', *bravo* 'bravo'

In a few cases they are morphologically related to regular words:

(55) *karar* 'men' (from *ka:r(l)* 'man'), *skiti(g)* 'dirty' (from *ski:t* 'shit')

Finnish loanwords and place names are normally pronounced with the CV́ syllables of the original:

(56) *poro* 'coffee grounds', *sisu* 'endurance', *kiva* 'fun, nice' (pl. *kivoga*, as if from a non-existent **kivog*, after Finnish partitive pl. *kivoja*)

This is not surprising, for practically all speakers of the Helsinki/Turku dialect speak Finnish too.

Other alternations which should be mentioned here for the sake of completeness are the following:

(57) a. /me/, /me:/, /meC/ *me Kickan* 'with Kickan', *ta: de mé:* 'take it along', *mém mej* 'wíth me', *me méj* 'with mé', *med dej* 'with you', etc. Similarly /påC/, /på/ *på Fö́lisö̀:n* 'on Fölisö', *sti:g på:* (**på*) 'come in', *pǻm mej* 'on me', *på méj* 'on mé'.

b. /i/, /i:/: *i* 'in', under the same conditions as (a) above.

c. But /ti/, /till/ *ti Fö́lisö̀:n* 'to Fölisö', *hjälpa till (*ti, *ti:)* 'help', *tíll mej* 'tó me', *ti méj* 'to mé'.

[31] There are also some interjections, but these of course are known to have special properties, and in fact can have stressed short vowels even in Sweden: *jahá* 'I see', *ahá* 'aha', *nå* 'nu', *tja* 'well', and *sí du* 'you see'.

This exhausts the cases where Helsinki Swedish has a three-way quantity contrast in stressed syllables. Elsewhere, it has the same two-way contrast as West Swedish.

Summary The grammatical restrictions on stressed light syllables become understandable in Stratal OT if we distinguish properly between the phonological constraints on stems and the phonological constraints on words. Stressed light syllables surface in those types of words that escape the stem-level constraints that prohibit them. (58) is a synopsis of the analysis:

(58)

	(41a)	(41b)	(41c)	(41d)
Underlying	/medan/	/hyvl/	/draj-/	/fotografi:/
Stems	—	[hývl]	—	fòtografí:
Words	[médan]	[hývel]	[dráji]	[fóto]

To recapitulate the main points of our Stratal OT analysis of Helsinki/Turku Swedish:

- Function words have stressed light syllables because they are not subject to the stem-level constraints.
- /CVCC/ words which become disyllabic CVCVC words though word-level epenthesis retain short vowels (case (41b)). At the stem level, they do not violate the prohibition on light stressed syllables. At the word level, the constraint is rendered inactive by dominant faithfulness constraints.
- The irregularly inflected verb forms (case (41c)) are formed by adding the regular inflected endings exceptionally to bound roots. Bound roots are not stems, and therefore do not undergo stem-level phonology. The outputs of the affixation process are words, and undergo only the word phonology. Accordingly, an underlying short vowel can surface in them.
- Finally, the truncation process responsible for case (41d) is applicable at the word level; this explains phonological properties of the truncatum, including its quantity.

Let us now integrate this analysis into the formal constraint system that we began to develop in earlier sections.

6.2.3 *The constraints*

The stem level We are now ready to incorporate the grammatical aspects of Helsinki Swedish quantity into our Stratal OT constraint system. In the *stem-level* phonology of Helsinki/Turku Swedish, stressed syllables must be bimoraic. That stressed syllables have exactly two moras (at this level) results from the two constraints in (59):

(59) a. $^*\mu\mu\mu$: No three-mora syllables.
b. STRESS-TO-WEIGHT (see (6c)).

Just as in West Swedish, light syllables are repaired by vowel lengthening, rather than by consonant gemination, and superheavy syllables are repaired by vowel shortening, rather than by degemination. Therefore, at the stem-level, the faithfulness constraints MAX-Cμ and DEP-Cμ must be outranked by the corresponding constraints for vocalic moras.

(60) a. MAX-Cμ: A consonantal mora in the input must correspond to a mora in the output.
b. DEP-Cμ: A consonantal mora in the output must correspond to a mora in the input.

The stem-level constraint system of the Helsinki dialect is:

(61) $^*\mu\mu\mu$, STR/W ((6c)) $\gg$ MAX-Cμ, DEP-Cμ

Tableau (62) shows for simple cases how the stem-level phonology makes all stressed syllables exactly two moras in length.

(62)

Stem Level		$^*\mu\mu\mu$	STR/W	MAX-Cμ	DEP-Cμ
Input: /mata/					
1a.	má.ta		*		
1b. ☞	máa.ta				
1c.	mát.ta				*
1d.	máat.ta	*			*
Input: /matta/					
2a.	má.ta		*	*	
2b.	máa.ta			*	
2c. ☞	mát.ta				
2d.	máat.ta	*			
Input: /ku/					
3a.	ku		*		
3b. ☞	kuu				

Underlying /maata/ will give the same output as /mata/. Similarly, underlying /maatta/ merges with /matta/. Thus, there are no stressed one-mora syllables or three-mora syllables at this level.

This illustrates Lexical Phonology's solution to the "duplication problem," which is also adopted in Stratal OT: the form of underlying representations is characterized by the same constraint system that governs *stem-level* alternations.

At the word level, *$\mu\mu\mu$ is dominated by FORTITION and by CODAGEMINATION. Because Helsinki's stricter stem-level phonology eliminates /CVVC/ syllables and /CV/ syllables (including /CVC words/), Fortition and Coda Gemination in this dialect will not produce violations of NONEUTRALIZATION in the cases considered so far. However the ranking of NONEUTRALIZATION in this dialect can be determined by other considerations. Recall that vowel length is distinctive only syllable-finally and before voiced consonants. In order to derive the length neutralization before voiceless consonants from Fortition, this constraint must outrank NONEUTRALIZATION. As the tableau makes clear, the distinction between the hypothetical inputs to the word level /medan/ and /meddan/ survives, whereas the inputs /deta/ and /detta/ merge into a single output as before. But the distinction can only be manifested in function words, where the CV inputs are available. In lexical words, they are eliminated at the stem level. In this way, the constraint system correctly reconstructs the fact that function words have an extra syllable type, but just before voiced consonants.

(63) Helsinki word-level ranking:
FORTITION ≫ CODAGEMINATION ≫ *$\mu\mu\mu$, NONEUTRALIZATION

(64)

Helsinki (W. L.)		FORTITION	CODAGEM	*$\mu\mu\mu$	NONEUTR
Input: /rista/ 'to carve'					
1a.	rís.ta	**	*		
1b. ☞	ríss.ta	*		*	
1c.	ríst.ta	*	*	*	
Input: /velja/ 'to choose'					
2a.	vél.ja		*		
2b. ☞	véll.ja			*	
2c.	vélj.ja		*	*	
Input: /ven.da/ 'to turn'					
3a.	vén.da		*		
3b. ☞	vénn.da			*	
3c.	vénd.da		*	*	

Input: /venta/ 'to wait'					
4a.	vén.ta	*	*		
4b.	vénn.ta	*		*	
4c. ☞	vént.ta		*	*	
Input: /riida/ 'to ride'					
5a. ☞	ríi.da				
5b.	ríii.da			*	
5c.	ríid.da			*	
Input: /riita/ 'to draw'					
6a.	ríi.ta	*			
6b.	ríii.ta	*		*	
6c. ☞	ríit.ta			*	
Input: /medan/ 'while'					
7a. ☞	mé.dan				
7b.	mée.dan				*
7c.	méd.dan				*
7d.	méed.dan			*	
7e.	médd.dan			*	
Input: /deta/					
8a.	dé.ta	*			
8b.	dée.ta	*			*
8c. ☞	dét.ta				*
8d.	déet.ta			*	
8e.	détt.ta			*	
Input: /nu/					
9a. ☞	nu				
9b.	nuu				*

In sum: voiced consonants have two special properties: they don't undergo Fortition, and they can be preceded by light open syllables. The constraint system (63) explains this intriguing correlation. It derives the basic Helsinki pattern where syllable weight is neutralized before voiceless consonants (merger of /-V́CV-/ and /-V́CCV-/ into [-V́CCV-]) precisely through Fortition. This implies the ranking Fortition $\gg$ NoNeutralization.

6.2.4 *Opacity*

The paradoxical anti-structure-preservation property of Fortition is related to another problem which the Stratal OT model also resolves. The process, which creates superheavy syllables consisting of a long vowel plus a geminate in the output, occurs only in those dialects which *prohibit* such superheavy syllables in *underlying* representations, and which shorten long vowels before geminates in derived words, as seen in (16b) and (18). The puzzle is that, in the output, the lexical restriction on superheavy syllables remains in force only for *voiced* geminates (*[ʃe:l.la], *[ro:d.de], see (1)). Before *voiceless* consonants, Fortition reintroduces the kinds of superheavy syllables that vowel shortening eliminates. This is a typical case of opaque constraint interaction, which Stratal OT claims is due to the serial relation between phonological levels.

As a simple illustration of how the opaque interaction between shortening and lengthening is explained by Stratal OT, consider *heta* [heet.ta] 'to be called' and *hette* [het.te] 'was called' in Helsinki Swedish. The stem is underlying /heet/ and the suffixes are /-a/ and /dde/. The derivations are as follows.

(65) a. /heet-a/ → [hee.ta] → [heet.ta] (word-level Fortition).
b. /heet-dde/ → [het.te] (*$\mu\mu\mu$ forces stem-level shortening) → [het.te]

In /heet-a/ → [heet.ta], word-level Fortition reintroduces the superheavy syllable structure at the word level that *$\mu\mu\mu$ eliminates in /heet-dde/ → [het.te] at the stem level. The paradox for parallelism is this. If CVVCCV is admissible, what forces vowel shortening [het.te]? Why do we not get just /heet-dde/ → [heet.te]? On the other hand, if CVVCCV is excluded, why /heet-a/ → [heet.ta]? Stratal OT's answer is that CVVCCV (and superheavy syllables in general) are admissible in words but excluded in stems. This instantiates Stratal OT's general solution to the problem of phonological opacity.

Parallel OT has two devices at its disposal for dealing with opacity: *Base/Output (Output/Output) constraints* (Benua 1997), and *Sympathy* (with or without Cumulativity, McCarthy 1999*b*, 1999*c*). Can either of these deal with these Swedish facts, in particular, with the shortening of the underlying vowel in /heet-te/ in the face of the admissible output [heet.ta]?

It appears that the answer is no. An Output/Output constraint would "borrow" the short vowel from somewhere else in the paradigm. But there is no such form, for the short vowel occurs *only* in the very cases that have to be explained, such as [het.te]. A Sympathy (or Cumulativity) constraint would "borrow" the short vowel from a failed candidate selected by some

FAITHFULNESS constraint. But there is no such candidate, simply because there is no more propitious shortening environment than the geminate that is seen in actual output itself. Simple though this case is, parallel OT seems to break down. The Swedish data clearly favor Stratal OT.

6.2.5 *Lexical diffusion*

I conclude with a brief historical remark. As noted in Reuter 1986, modern Helsinki Swedish is practically unchanged as far as quantity is concerned since Bergroth 1922 and 1917/1928 (and, apparently, since Pipping 1892–7, which however I have not seen). The stability is remarkable, considering that short light syllables have been stigmatized in schools at least since the publication of Bergroth's orthoepic handbook in 1917, and very likely even earlier. The main changes are that a number of lexical items whose short stressed vowels are unpredictable on the present account have been regularized.

(66) a. Bergroth 1922: *juni* 'June', *juli* 'July', *huvu* 'head', *ströming* 'herring', *fräken* 'freckle', *stuli* 'stolen', *svuri* 'sworn', *skuri* 'cut'

b. Reuter 1986: *ju:ni*, *ju:li*, *huvvu* (but *Hufvudstadsbladet* [hʉ́vʉstasblàade(t)] 'The Capital Paper', a newspaper), *strömming*, *stu:li*, *svu:ri*, *sku:ri*, *frä:ken*

The short-vowel forms in (66a) are outright exceptions on the present account, analogous to those in (54)–(56). The regularized forms were normal by *ca.* 1950. In general, the 1986 situation reported by Reuter is identical to the one I recall from that time; the only recent change I find there is that a few forms, such as *sku vila* 'would like to' and *sku boda* 'should', which I think earlier were fairly standard, are now said to be used only by lower-class speakers.

In diachronic perspective, the development of Helsinki Swedish light stressed syllables constitutes a typical case of lexical diffusion. The theory of lexical diffusion proposed in Kiparsky (1995) (adapted to Stratal OT in the obvious way) explains the site and direction of the change as the elimination of arbitrary complexity from the lexicon, with resulting reversion to the unmarked state. The historical record shows that precisely those words which the present analysis characterizes as exceptions that require marking in the lexicon are being slowly eroded on an item-by-item basis, and that precisely in those word classes where vowel shortness is regular according to the present theory, it has managed to resist the uplifted fingers of pedagogues for the better part of a century.

6.3 Conclusion

The distribution of syllable weight in Fenno-Swedish dialects is governed by an anti-neutralization constraint and by the interaction of distinct constraints on stems and words. Both were shown to support a stratal version of OT phonology against parallel OT.

7

SPE Extensions: Conditions on Representations and Defect-Driven Rules[1]

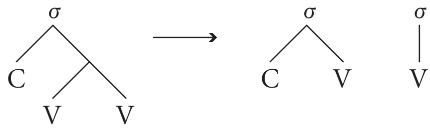
JOHN FRAMPTON

In this chapter I propose a new formal theory of iterative rules and demonstrate the advantage of analyzing various prosodic phenomena in terms of the rule format which is introduced. Noteworthy is the extensive use of constraints and well-formedness conditions in the formulation of individual rules. The title "SPE Extensions" was chosen to highlight the fact that, in spite of the fact that extensive use is made of conditions on representations, the proposed theory adheres closely to the traditional framework of sequential rule application proposed in *The Sound Pattern of English* (Chomsky and Halle 1968). In the proposed format, the application of iterative rules is driven by defects (along some dimension) in the structure they operate on. This application can be extensively controlled by systems of derivational constraints. The proposed format relies heavily on the schema interpretation and expansion mechanisms proposed in SPE.

The particular set of problems which motivated the development of this theory of iterative rules concerns instances of syllable restructuring in order to accommodate the demands of foot structure. In various languages, CVV syllables split into CV.V syllable pairs in certain environments to accommodate the requirements of foot formation. This interaction between footing and syllable structure is difficult to analyze in a theory in which syllable structure is modeled on syntactic $\overline{\text{X}}$-theory (Levin 1985). If the structural change CVV $\rightarrow$ CV.V is something along the lines of (1) and feet are constructed on an unrelated

[1] I am grateful to Morris Halle for his guidance, support, and contagious curiosity about how the language faculty does what it does. Thanks also to Sylvain Bromberger, Sam Gutmann, Wayne O'Neil, Eric Raimy, and Moira Yip, as well as audiences at MIT and at the Phonology 2000 Conference, for helpful comments. This chapter is a rewriting of part of a longer paper which circulated as "SPE Extensions" in 1999.

tier, it stretches plausibility to imagine that the two operations take place as two aspects of one operation.

(1) 

This chapter is devoted to working out a theory of syllable structure which makes syllable restructuring a simple operation and a theory of footing which ties it directly to syllable structure. In the framework which is developed, the accommodation of syllable structure to footing will be virtually transparent. Defect-driven iterative rules will be crucial in both syllabification and footing. Section 7.1 develops a formal theory of defect-driven iterative rules, using a variety of simple footing rules (Garawa, Southern Paiute, Cayuvava, Hawaiian) to illustrate the idea. Section 7.2 uses defect-driven iterative rules to construct a tier, called the *cluster tier*, roughly akin to a mora tier. The cluster tier and associations with the timing tier contain all the information that is usually thought to come from syllable structure. The iterative syllabification rule is worked out in detail for Icelandic and Imdlawn Tashlhiyt Berber. Section 7.3 brings foot structure and syllable structure together by proposing that footing is carried out by inserting delimiters into the cluster tier. Cairene Arabic is used to illustrate the effect of Syllable Integrity (*Split-× in my reformulation) on delimiter insertion. Section 7.4 shows that syllable splitting under the demands of footing can be realized as elementary autosegmental delinking and relinking. The phenomenon is illustrated for Fijian, Tongan, Southern Paiute, and Gothic (Sievers' Law).

7.1 Defect-driven iterative rules

Kisseberth (1970) first discussed rules which apply only when the input representation violates some constraint and the output does not. The core idea was further developed and articulated by Sommerstein (1974). The intention in this chapter is to make this the organizing principle of iterative rule application. The idea is that iterative rules are driven by defects in the structure they apply to, in the sense that they apply if and only if they can remove a defect, iterating until they can no longer apply. Application terminates either because all defects have been removed or because the rule provides insufficient resources to remove any remaining defects.

In order to illustrate the idea, we begin with left to right footing, adopting Idsardi's view of footing (Idsardi 1992; Halle and Idsardi 1995). Footing is delimiter insertion, one delimiter at a time. For the moment, I abstract away from what units are footed and call the units *stressable elements*. The traditional account of left to right iterative binary footing is that the rule (2a) applies iteratively, producing derivations like that in (2b).

(2) a. Ø → ⟩ / ○ ○ ___ (left to right)

b. ○ ○ ○ ○ ○ → ○ ○⟩○ ○ ○ → ○ ○⟩○ ○⟩○

This can be recast as a defect-driven rule by imposing the target condition (3):

(3) / ○___ ⇒ / ___⟩

Condition (3) requires that a stressable element which immediately follows another stressable element (i.e. it is in the context ○___) must be immediately followed by a right foot delimiter (i.e. it must be in the context ___⟩). Note that all the stressable elements in the final representation in (2b) satisfy this condition. Left to right footing will be recast as an iterative rule which brings the stressable elements progressively into line with Condition (3), which will be called ⟩-Delimited. A stressable element which is does not satisfy ⟩-Delimited is defective (with respect to left to right iterative footing).

Suppose the initial representation is:

○ *○ *○ *○ *○

The defective stressable elements have been annotated with an asterisk. Demanding that the defects be progressively removed, from left to right, is not sufficient to ensure the desired derivation. Along with the desired (4a), many undesired derivations are also possible;

(4) a. ○ *○ *○ *○ *○ → ○ ○⟩○ *○ *○ → ○ ○⟩○ ○⟩○

b. ○ *○ *○ *○ *○ → ○ ○⟩○ *○ *○ → ○ ○⟩○⟩○ *○ → ○ ○⟩○⟩○ ○⟩

c. ○ *○ *○ *○ *○ → ○⟩○ *○ *○ *○ → ○⟩○⟩○ *○ *○

→ ○⟩○⟩○⟩○ *○ → ○⟩○⟩○⟩○⟩○

Some factor is needed to prevent delimiters from being inserted too close together. A constraint against unary feet (which I call *Uny) plays an important role in many footing systems. I assume it is in force here, so that derivations (4b) and (4c) are excluded and (4a) is forced, as desired.

What is required is a rule format which can make all of this explicit. The general format of a *Defect-Driven Rule* is:

(5) $\underbrace{\text{Type ; Condition ; Order}}_{\text{Preamble}}$ :: $\underbrace{\text{Rule(s) ; Constraint Set}}_{\text{Body}}$

The condition in the preamble (on elements of the specified type) is called the *target condition*. Order ranks the violations and determines the order in which the rule attempts to remove them. The rules in the body are called *repair rules*. Constraints is a list of constraints on repair rule application. The metaphor is that violations of the target condition are defects which the rules repair.

The simple left to right binary footing rule discussed above, written in this format, is:

(6) Stressable Element ; ⟩-Delimited ; Left :: Ø → ⟩ ; {*Uny}

Now consider a slightly more complex footing rule of this type, this one for right to left footing. The right to left footing condition / __ ○ ⇒/ ⟨ __ , called ⟨-Delimited, is the mirror image of the left to right footing condition.

(7) Stressable Element ; ⟨-Delimited ; Right :: $\begin{bmatrix} \emptyset \rightarrow \langle \\ \emptyset \rightarrow \rangle \end{bmatrix}$; {*Uny, *#o⟨}

There are two derivational constraints. *Uny, as above, and *#o⟨, which bars orphaned stressable elements at the left edge. As a notational convenience, # indicates the left edge and % the right edge. The effect of *#o⟨ is to force foot alignment with the left edge. Two rules are available for removing defects. The rule schema is interpreted in the usual SPE fashion, the highest-ranked rule which can apply, does apply. The highest-ranked defect is first chosen, then the highest-ranked rule which can remove this defect applies.

This gives the derivations in (8). Arrow subscripts are used to provide reference points for the discussion that follows.

(8) a. *○ *○ ○ $\rightarrow_1$ *○ ○⟩○ → ⟨○ ○⟩○

b. *○ *○ *○ ○ → *○ ○⟨○ ○ → ⟨○ ○⟨○ ○

c. *○ *○ *○ *○ ○ → *○ *○ ○⟨○ ○ $\rightarrow_2$ *○ ○⟩○⟨○ ○ → ⟨○ ○⟩○⟨○ ○

In steps 1 and 2, the highest-ranked rule, Ø → ⟨, cannot remove the highest-ranked (rightmost) defect because of *Uny and *#o⟨. The next highest-ranked rule, Ø → ⟩, therefore applies to remove the rightmost defect, in the only way that it can. Inserting ⟩ to the left of the rightmost defect would not remove the defect. This is the footing system in Garawa (see Halle and Idsardi 1995: 422, for discussion and examples).

We now consider another example of this type which illustrates how the derivational constraints, the preferences of the rule schema, and repair rule preference, interact. Again, right to left footing of strings of stressable elements is the issue. The edge-marking rule (9.EM) applies prior to the defect-driven iterative footing rule (9.IF).

(9) EM : $\emptyset \rightarrow \rangle \,/\, \_\_ \circ \%$

IF : Stressable Element ; ⟨-Delimited ; Right :: $\begin{bmatrix} \emptyset \rightarrow \rangle \\ \emptyset \rightarrow \langle \end{bmatrix}$;

$\{{}^{*}\text{Uny}, {}^{*}\#\langle\}$

For two- and three-element strings, edge marking produces:

(10) a. $\circ \rangle \circ$ b. $^{*}\circ \;\; \circ \rangle \circ$

The iterative footing rule cannot remove the defect in (10b) because of the two derivational constraints. It therefore does not apply and the iteration terminates. This illustrates two important points. It is not uncommon that a violation cannot be removed with the given repair rule resources. This is the case in (10b). The second point is that although the initial element in (10b) remains defective, it is footed. In fact, if stress is trochaic (assigned to the leftmost stressable element of feet), the defective element gets main stress in (10b). The view of the conditions which drive iterative footing adopted here is very different from the view taken by Prosodic Morphology (see McCarthy and Prince (1995b), and the many references cited there), which takes the driving force for the organization of prosodic structure to be membership in prosodic categories of one type or another.

A four-element string produces the derivation:

(11) $\circ \;\; \circ \;\; \circ \;\; \circ \;\; \overset{\text{EM}}{\rightarrow} \;\; {}^{*}\circ \; {}^{*}\circ \;\; \circ \rangle \circ \;\; \overset{\text{IF}}{\rightarrow}_{1} \;\; \circ \langle \circ \;\; \circ \rangle \circ$

No defective elements are marked in the initial representation because "defective" (with respect to the target condition of the iterative footing rule) has no meaning outside of application of the iterative footing rule. The target condition $/ \_\_ \circ \Rightarrow / \langle \_\_$ is not a general condition on representations, but a specific condition used to organize right to left iterative footing. In this, it is more akin to the structural description of a rule than to a constraint on representations. Note at step 1 that the most highly ranked rule cannot apply to remove the rightmost violation. Insertion of ⟩ to the right of the rightmost defective element would violate *Uny, while insertion to its left would not remove the violation.

A derivation starting from a longer string of stressable elements is given below:

(12) $\circ\ \circ\ \circ\ \circ\ \circ\ \circ\ \circ \overset{\text{EM}}{\to} {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ \ \circ\rangle\circ \overset{\text{IF}}{\to} {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ \ \circ\langle\circ\ \ \circ\rangle\circ$
$\overset{\text{IF}}{\to} {}^*\!\circ\ {}^*\!\circ\ \ \circ\rangle\circ\langle\circ\ \ \circ\rangle\circ \overset{\text{IF}}{\to} \circ\langle\circ\ \ \circ\rangle\circ\langle\circ\ \ \circ\rangle\circ$

If stress is assigned foot left and main stress is assigned word-right, the footing rules (9) produce the ternary stress pattern of Cayuvava. See Hayes (1995: 309). It is instructive to compare the footing pattern produced by (9) with the footing pattern produced by the almost identical (13), in which the only change is that the rankings of the two repair operations in the iterative footing rule have been reversed.

(13)
EM : $\varnothing \to \rangle \ / \ \_\_\ \circ\ \%$
IF : Stressable Element ; ⟨-Delimited ; Right :: $\begin{bmatrix} \varnothing \to \langle \\ \varnothing \to \rangle \end{bmatrix}$; $\{{}^*\text{Uny}, {}^*\#\langle\}$

The footing rules (13) yield binary footing rather than ternary footing.

(14) $\circ\ \circ\ \circ\ \circ\ \circ\ \circ\ \circ\ \circ \overset{\text{EM}}{\to} {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ \ \circ\rangle\circ$
$\overset{\text{IF}}{\to} {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ {}^*\!\circ\ \ \circ\langle\circ\ \ \circ\rangle\circ$
$\overset{\text{IF}}{\to} {}^*\!\circ\ {}^*\!\circ\ \ \circ\langle\circ\ \ \circ\langle\circ\ \ \circ\rangle\circ$
$\overset{\text{IF}}{\to} \circ\langle\circ\ \ \circ\langle\circ\ \ \circ\langle\circ\ \ \circ\rangle\circ$

In fact, (13.IF) never needs to resort to the less highly ranked $\varnothing \to \rangle$, so that (13.IF) is effectively

Stressable Element ; ⟨-Delimited ; Right :: $\varnothing \to \langle$; $\{{}^*\text{Uny}, {}^*\#\langle\}$

A small change in the ranking of the repair rules switches footing from ternary to binary.[2]

7.1.1 *Discretionary constraints*

Up to this point, we have assumed that the constraint set consists of only *strict constraints* on rule application. We now admit a new class of constraints, called *discretionary constraints*. In removing a given defect, discretionary constraints can be violated, but only as a last resort. Multiple (ranked) discretionary

[2] Plausibly, the relative rarity of ternary footing can be attributed to the fact that $\varnothing \to \langle$ is the natural repair rule for satisfying ⟨-Delimited.

constraints require a discussion of constraint-set schemata and expansion, but for a single discretionary constraint, the idea of violation only as a last resort will suffice. This will be sufficient for the examples in this chapter.

We can see how this works by considering the Southern Paiute iterative footing rule (15). Discretionary constraints come after the || symbol.

(15) Stressable Element ; ⟩-Delimited ; Left :: Ø → ⟩ ; $\{$ *⟩% || *Uny $\}$

The effect of discretionary *Uny is that the creation of unary feet is not absolutely excluded, but is legitimate only if a particular defect cannot be otherwise removed.[3]

A few representative derivations follow. There is no edge marking.

(16) a. ○ *○ $\rightarrow_1$ ○⟩○

b. ○ *○ *○ → ○ ○⟩○

c. ○ *○ *○ *○ → ○ ○⟩○ *○ $\rightarrow_2$ ○ ○⟩○⟩○

d. ○ *○ *○ *○ *○ → ○ ○⟩○ *○ *○ → ○ ○⟩○ ○⟩○

A violation of discretionary *Uny is forced at steps 1 and 2 because the alternative option for removing the defect is blocked by non-discretionary *⟩%. With foot stress right (iambic) and word stress right, this is the stress system of Southern Paiute. Main stress is always penultimate.

As a second example, consider the Hawaiian footing system:

(17) EM: Ø → ⟩ / ___ %

IF: Stressable Element ; ⟨-Delimited ; Right :: $\begin{bmatrix} Ø \rightarrow \langle \\ Ø \rightarrow \rangle \end{bmatrix}$;

$\{$ *Uny || *#○⟨ $\}$

Some representative derivations follow:

(18) a. *○ ○⟩ → ⟨○ ○⟩

b. *○ *○ ○⟩ $\rightarrow_1$ ○⟨○ ○⟩

c. *○ *○ *○ ○⟩ → *○ ○⟨○ ○⟩ → ⟨○ ○⟨○ ○⟩

d. *○ *○ *○ *○ ○⟩ → *○ *○ ○⟨○ ○⟩ $\rightarrow_2$ *○ ○⟩○⟨○ ○⟩

→ ⟨○ ○⟩○⟨○ ○⟩

[3] It is likely that discretionary *Uny is universal in iterative footing, so that it does not have to be explicitly mentioned in (15).

At step 1, there is a violation of discretionary *#o⟨, but the alternative ways of removing the rightmost defect violate non-discretionary *Uny. At step 2, the highest-ranked rule, $\emptyset \rightarrow \langle$, is not used because it would require violation of either non-discretionary *Uny or discretionary *#o⟨. The less highly ranked repair rule $\emptyset \rightarrow \rangle$ therefore applies.

If stress is assigned foot-left and word-left, this is the stress system of Hawaiian. Main stress is initial, except if there are three stressable elements, in which case it is medial.

7.1.2 *Defectiveness is local to the rule which mentions it*

The idea that the most fruitful way to view the structural conditions which trigger rule application, at least certain instances of rule application, is in terms of defects in the structure which rule application can remove, has received ongoing attention since the early work of Kisseberth and Sommerstein (see Singh 1987; Yip 1988; Paradis 1988a, b; Goldsmith 1990; Calabrese 1995). Lacharité and Paradis (1993) give a useful survey.[4] In all of this work, the concern is for deviation from high-level conditions on phonological structure. The tendency has been to assimilate the idea of "defect" to the idea of violation of an output condition or a surface phonotactic. Only Calabrese (see p. 411, for example) allows well-formedness conditions which trigger repair at some stages of the derivation but are inactive at other stages.[5] This tendency to view defects as deviations from surface phonotactics fed Optimality Theory's preoccupation with surface well-formedness.

The conception of "defect" proposed here is much less closely tied to the idea of well-formedness condition imposed on surface structure. It is more akin to the notion of a "structural description" in SPE terms. A particular notion of defect need only be relevant to the iterative rule in which it appears. There are reasons both from tendencies towards formal simplicity and from learning theory for well-formedness conditions which appear at one place in the grammar to appear at other places as well, perhaps even at the surface, but these reasons are external to the formal constraints on possible grammars and to the online computation which the grammar specifies.

The separation of well-formedness conditions which drive defect-driven iterative rules from surface phonotactics is illustrated, for example, by the condition that drives left to right iterative footing: $/\ \circ \_\_ \Rightarrow /\ \_\_ \rangle$. One of the major advances of prosodic theory was Liberman's (1975) discovery that stress placement depended on the abstract computational device of foot

4 Unaccountably, the work of Calabrese is not mentioned.

5 In more recent work, Calabrese supposes that constraints are associated with particular strata.

formation. The constraint $/ \circ \underline{\quad} \Rightarrow / \underline{\quad}\rangle$ does not even make sense as a phonotactic, since foot delimiters have no immediate phonetic correlate. The computation of foot structure has an effect on surface structure principally because it feeds the computation of stress, accent, and tone, which is what is most clearly visible at the surface. The abstractness of foot structure, and therefore of the target condition which drives it, is made particularly clear in a language such as Cyrenacian Bedouin Arabic in which syncope rules applying after foot formation cause widespread deviance from the target condition which organizes foot construction. Syncope completely obscures the origin of the foot structure which produces the observable surface stress patterns. (See Frampton (1999), for an analysis in the present framework.)

It is also worth reminding the reader that defect-driven iterative rules are not necessarily successful in removing all the defects which drive their application. Defects can remain after rule application. There is no guarantee that the array of rules (repair rules) which a defect-driven iterative rule allocates to defect removal is sufficient for the job. The rule operates by doing the best it can, and then moving on. Note also that the notion of "repair rule" used here, like the notion of defect itself, is particular to the defect-driven iterative rule in question. Indeed, we will see in Section 7.2 that the array of repair rules which the iterative syllabification rule specifies is a major determinant of syllable structure.

7.2 Autosegmental syllabification

Itō's (1986) templatic theory of syllabification made significant advances over earlier theories by integrating well-formedness conditions on syllable structure into a syllabification algorithm built around directional template matching. But Dell and Elmedlaoui's (1985) study of syllabification in Imdlawn Tashlhiyt Berber (henceforth ITB) showed that templatic matching was inadequate and proposed a two-stage theory of syllabification for ITB in which core syllables were constructed first, then full syllables built around them. The key innovation was abandoning directional syllabification (left to right, or right to left) and making the order in which syllables are built dependent on the sonority of the phonemes which were to be syllabified.

It is the intention of this section to develop Dell and Elmedlaoui's idea into a defect-driven iterative rule that accomplishes full syllabification, not just core syllabification, and not only in ITB, but (with language-particular variation) quite generally. The proposed syllable structure will be autosegmental, which is very natural from the standpoint of iterative rule application. It has the virtue that the kinds of syllable reorganization that occur in response to footing are elementary delinking and relinking processes familiar from autosegmental phonology. We will return to investigate the interaction of syllable structure and foot structure in Section 7.4.

Syllabification is a process of grouping timing slots. I propose that it is implemented by the construction of a tier, the *cluster tier*, and associating elements on this tier, *clusters* (denoted by ω below), with timing slots. There are two elementary cluster formation rules, given in (19). They should be understood as operations on $\times$.

(19) Form Doublet $\times' \quad \times \rightarrow \overset{\omega}{\widehat{\times' \quad \times}}$

Form Singlet $\times \rightarrow \overset{\omega}{\underset{\times}{|}}$

The asymmetry of Form Doublet is important. The rule operates on the timing slot $\times$ and *builds a cluster to the left.*

Suppose $\mathcal{L}$ is a simple (hypothetical) language in which vowels cannot be onsets and non-vowels cannot be nuclei. Consider the defect-driven iterative rule:

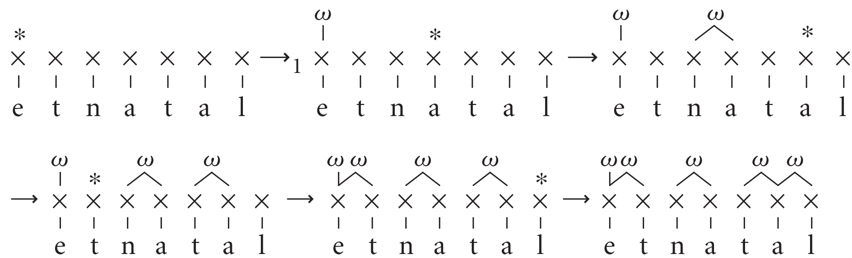
(20) $$\text{Timing Slot} \;\; ; \; \text{Clustered} \; ; \begin{pmatrix}\text{Vowel}\\ \text{Left}\end{pmatrix} :: \begin{bmatrix}\text{Form Doublet}\\ \text{Form Singlet}\end{bmatrix}$$

$x > y$ with respect to the order (Vowel Left) is x is a vowel and y is not, or if both x and y are vowels or non-vowels, but x is to the left of y.[6] The order component of the footing rules considered in the last section dealt with defects in a directional manner, left to right or right to left. The order here is not primarily directional. The primary determinant of order is the vowel/non-vowel status of a timing slot. Vowels are targeted for repair first. Directionality is only relevant for distinguishing between timing slots which are not distinguished by Vowel.

The rule (20) produces, for example, the derivation (21). To aid the reader, the highest-ranked defect in each representation is marked with an asterisk.

[6] In general, predicates can be interpreted as precedence relations. If P is a predicate, then we say $x > y$ with respect to P if $P(x)$ and not $P(y)$. We also say $x = y$ with respect to P if not $x > y$ with respect to P and not $y > x$ with respect to P. Predicates over phonemes can be interpreted as predicates over timing slots via association. The complication of timing slots associated with multiple phonemes will be ignored here, since it is irrelevant to what follows. A list of precedence relations is itself interpreted as a precedence relation by saying $x > y$ with respect to $(P_1 P_2 \ldots P_n)$ if $x > y$ with respect to P_1, or $x = y$ with respect to P_1 and $x > y$ with respect to $(P_2 \ldots P_n)$. The recursion is terminated by interpreting (P) as P.

(21)



At step 1, the highest-ranked rule (Form Doublet) cannot remove the highest-ranked defect, but the next highest ranked rule (Form Singlet) can.

In spite of its unorthodox form, the representation produced in (21) has all the relevant information about "syllable structure" that more orthodox representations have.

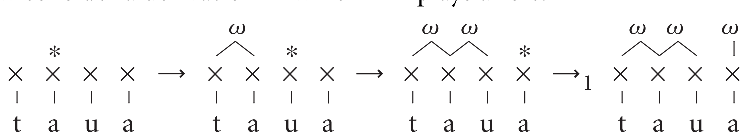
We define a *syllable* to be a maximal connected set of clusters. The leftmost (perhaps only) cluster of a syllable is called its *onset cluster*. Clusters which are not onset clusters will be called *coda clusters*. Both clusters and syllables have heads. I assume that: (1) every syllable head is a cluster head, and vice versa; and (2) clusters are rightheaded, if possible. It is easy to see that this implies that:

(22)
1. a syllable can contain at most two clusters;
2. onset clusters are rightheaded; and
3. coda clusters are left-headed.

The constraint against syllables with more than two clusters will be called *Tri (i.e. no triclusters).

Now consider a derivation in which *Tri plays a role.

(23)



Application of Form Doublet is blocked by *Tri at step 1. The lower-ranked Form Singlet is therefore used to remove the defect.

The notion of "cluster" introduced here is related to the notion of "demi-syllable" used by Clements (1990) in his study of syllable sonority profiles:[7]

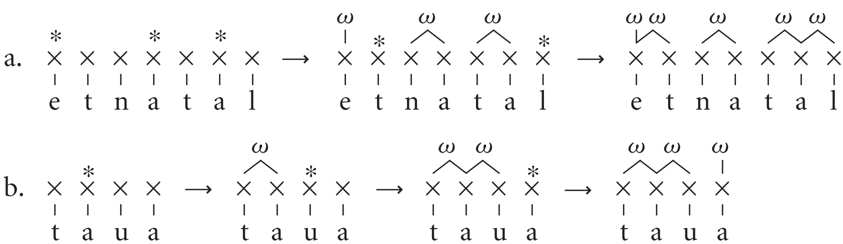
[7] Clements says: "The demisyllable was first introduced as a linguistic unit in the acoustic and phonological studies of Fujimura and his colleagues (see, e.g., Fujimura *et al.* 1977), but has not previously received explicit phonological justification." I have not been able to consult the papers which Clements refers to.

"Let us refer to syllable halves—overlapping portions of the syllable bounded at one end by the peak—as *demisyllables*." The new terminology *cluster* is justified because in the theory proposed here cluster is the primitive notion and syllable the derived notion. The rule (20) is not a syllabification rule, properly speaking, but a clustering rule.

- *Simultaneous application at multiple local maxima*

In the five-step derivation (21) above, the global maximum of (Vowel Left) was determined at each step. Suppose instead that the body of the iterative rule applies to all local maxima of (Vowel Left) simultaneously. An unclustered timing slot $\times$ is a *local maximum* if there is no other unclustered timing slot $\times'$ adjacent to $\times$ with $\times' > \times$ with respect to (Vowel Left). The derivations of (21) and (23) in this mode of application of iterative footing are given below, with the local maxima at each step marked with an asterisk.

(24)



The output is identical to the mode of application in which a global maximum is found at each step. In the remaining syllabification examples in this section, simultaneous multiple application at local maxima gives the same output as strictly sequential application at global maxima. Since a formal proof of this equivalence depends upon the clustering rules which apply, their order, and the kinds of constraints allowed, a detailed treatment is beyond what I can do in this chapter. Nevertheless, for the sake of minimizing the number of steps that need to be illustrated, I will assume simultaneous application at multiple local maxima throughout. The reader can check in any doubtful case that this mode of application gives the expected results.

- *The core structural inventory (CSI)*

Typically, only a subset of the full phoneme inventory is eligible to be syllabified as syllable nuclei. Of these, some can also occur as onsets. This distribution has a major impact on the iterative syllabification rule. It is the major determinant of the order in which defects are targeted for repair. I suppose, as a parametric choice, that each language establishes a *core structural* *inventory* (CSI) by distinguishing a class of *strongly nuclear* phonemes, which can be nuclei but cannot be onsets, and a class of *semi-nuclear* phonemes, which can be either onsets and/or nuclei. Remaining phonemes (*non-nuclear* phonemes) cannot be nuclei.

Given the CSI, we define the precedence relation $\gg$, "more nuclear than," between phonemes. $\alpha \gg \beta$ if α is strongly nuclear and β is not, or if α is semi-nuclear and β is non-nuclear, or if α and β are both semi-nuclear, but α is more sonorous than β. Iterative syllabification targets local maxima of $\gg$, with directionality deciding the loci of syllable-building in case adjacent timing slots are both local maxima of $\gg$.

As an example, consider a language with the phoneme inventory:

$$\left\{a, u, i, r, l, m, n, \text{ non-sonorants}\right\}$$

Suppose the CSI designates $\{a, u, i\}$ as strongly-nuclear and $\{r, l, m, n\}$ as semi-nuclear. The relation $\gg$ then partitions the phonemes as follows:

$$a, u, i \gg r, l \gg m, n \gg \text{non-sonorants}$$

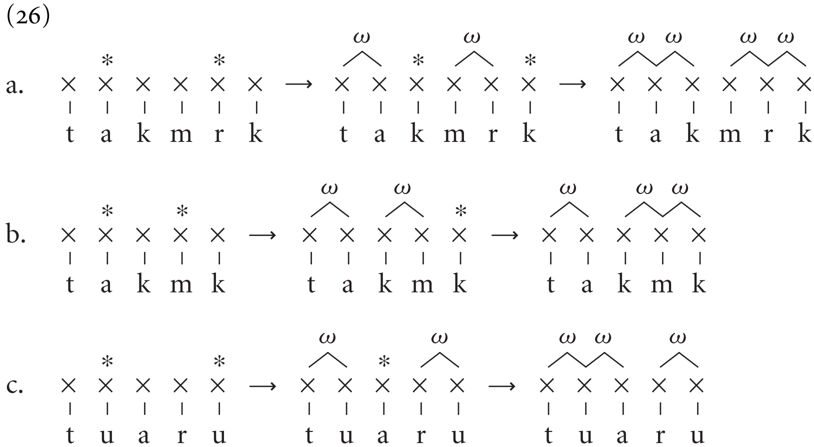
The partition of the semi-nuclear phonemes by sonority is superimposed on the tripartite partition into nuclear, semi-nuclear, and non-nuclear phonemes.

Now consider (25), a variant of (20), in which Vowel is replaced by the more fine-grained relation $\gg$.

(25) $$\text{Timing Slot} \;;\; \text{Clustered} \;;\; \begin{pmatrix} \gg \\ \text{Left} \end{pmatrix} :: \begin{bmatrix} \text{Form Doublet} \\ \text{Form Singlet} \end{bmatrix}$$

If no conditions on syllable shape intervene, we would expect:

(26)



Note that the semi-nuclear nasal *m* is syllabified as an onset when followed by the more sonorant semi-nuclear liquid *r*, in (26a), but as a nucleus in (26b). In (26c), on the other hand, *r* is syllabified as an onset, since it is followed by the strongly nuclear *u*. Note also in (26c) that the initial local maxima does not fall on *a*, even though it is more sonorant than *u*. The relation $\gg$ does not distinguish between strongly nuclear phonemes, so *u* is chosen over *a* on the basis of Left.

- *The default defect-driven iterative syllabification rule*

These considerations lead to the proposal that, given the CSI (which establishes the $\gg$ relation), the default choice for iterative syllabification is (25). The language learner uses the default rule as the starting point and will modify it to suit the particularities (over and above the CSI) of the language. The array of repair rules is often augmented in particular languages by rules which form complex onsets and complex codas, epenthesize onsets or nuclei, and delete otherwise unsyllabifiable elements. Language-particular conditions (a constraint on possible codas, for example) can be added to the constraints on repair operations. There is also variation in the directionality component of the order in which defects are targeted.

A thorough exploration of the proposal that syllable structure is autosegmental is beyond the scope of this chapter. The discussion will be limited to two languages, ITB and Icelandic. By considering two quite different languages, the hope is that universal autosegmental syllable structure will be made at least plausible.

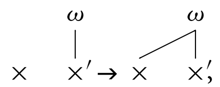
7.2.1 *Icelandic*

Syllabification in Icelandic is treated in some detail in Itō (1985), from which the examples in this section are taken. The intention here is not to break any new ground, simply to provide a good illustration of iterative autosegmental syllabification for a well-known language and to show how various syllabification phenomena are handled.

The CSI designates the vowels as strongly nuclear and does not designate any semi-nuclear phonemes. The relation $\gg$ is therefore the relation Vowel, viewed as a precedence relation between timing slots. Complex onsets are permitted, but they must have strictly increasing sonority. Certain consonant sequences cannot be broken up between coda and onset. Say that a consonant sequence is an *exceptional consonant sequence* if the first consonant is one of $\{p, t, k, s\}$ and the second is one of $\{r, j, v\}$. The phonemes represented as *j* and *v* in Icelandic are sonorants, so exceptional consonant sequences have increasing sonority. Lastly, unsyllabified elements are deleted.

The default iterative syllabification rule (25) is modified in several ways.

1. A rule Adjoin Onset is included in the array of repair rules, and constrained so that it applies only if the sonorities are appropriate.

(27) Adjoin Onset:  if $\times' > \times$ wrt Sonority

The rule is considered to be an operation on × and the autosegmental diagram is interpreted non-exclusively (i.e. the requirement is only that the given association exists, not that it is the exclusive association of the items it associates).

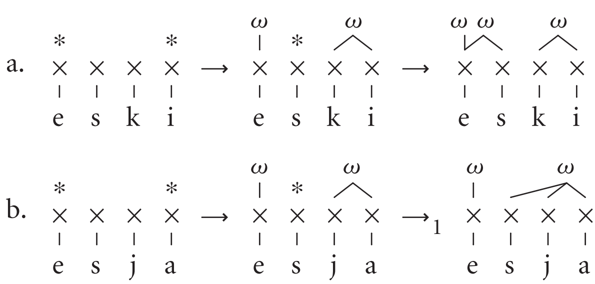
2. Form Doublet is restricted so that it does not apply to a timing slot linked to the first consonant of an exceptional consonant sequence.[8]
3. A rule Delete Timing Slot ($\times \rightarrow \emptyset$) is added to the list of repair rules as a least favored option.
4. The CSI is slightly relaxed word-finally: *r* can head a word-final monocluster.

Given these modifications, the iterative syllabification rule is:

$$\text{(28)}\quad \text{TS(x)} \;;\; \text{Clustered} \;;\; \begin{pmatrix}\text{Vowel}\\ \text{Left}\end{pmatrix} :: \begin{bmatrix}\text{Form Doublet}\\ \text{Form Singlet}\\ \text{Adjoin Onset}\\ \text{Delete Stray}\end{bmatrix} ;\; \{\text{CSI}, {}^{*}\text{Tri}\}$$

First, two examples illustrate the behavior of exceptional consonant sequences.

(29)

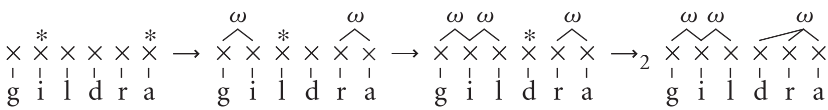


sj is an exceptional consonant sequence, so Form Doublet cannot apply at step 1. Form Singlet is blocked by the CSI, which disallows consonantal nuclei. Adjoin Onset therefore applies.

[8] The CSI prevents it from applying to the second consonant of an exceptional consonant sequence.

The example below shows that multiple onsets are not restricted to exceptional consonant sequences.

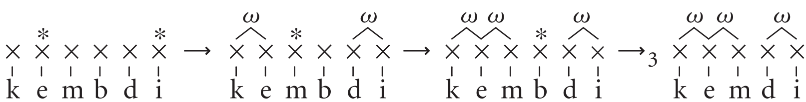
(30)



Form Doublet is blocked at step 2 by *Tri. Form Singlet is blocked by the CSI. So Adjoin Onset applies.

Timing slot deletion applies in (31):

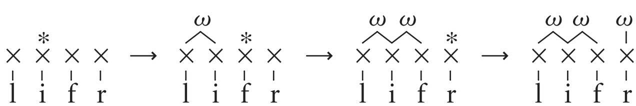
(31)



Form Doublet and Form Singlet are blocked at step 3 as in (30). But here, Adjoin Onset cannot apply because of the restriction to increasing onset sonorities. Therefore, the lowest-ranked rule, Delete Timing Slot, applies.

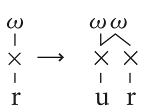
The effect of the relaxation of CSI to allow word-final monocluster *r* shows up in (32).

(32)



This is the result of iterative syllabification, which I suppose applies cyclically. Postcyclically, resyllabification applies. One of the defects which resyllabification is aimed at (in Icelandic and typically) is monoclusters. *u*-epenthesis applies in Icelandic to eliminate the final monocluster.

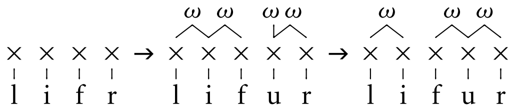
(33)



The monocluster which is targeted is eliminated, but a new monocluster is created. This is characteristic of defect-driven iterative rules. Elimination of one defect can create another one. Vowel harmony rules are a clear example; alteration to produce harmony to the left can create disharmony to the right. Iteration eventually removes all the defects, with directionality preventing endless looping. The new monocluster here is eliminated, as is common in resyllabification, by stealing the coda of the preceding syllable to provide an onset.

Resyllabification therefore yields:

(34)

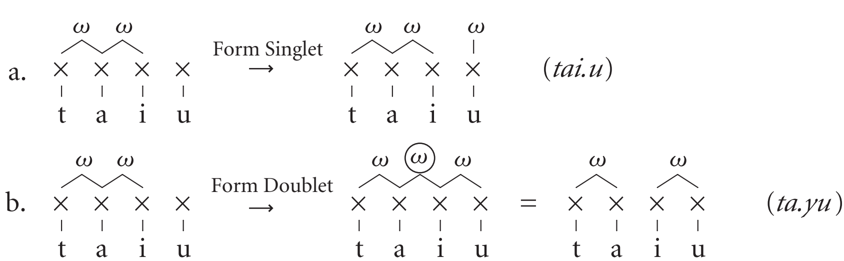


Itō obtains this result by supposing that there is a special template association principle for *r* which stipulates that it can associate with the coda position in a syllable with an empty head. It is crucial to Itō's, account that the nucleus not be realized as *u* until the postcyclic morphology, since this epenthetic *u* fails to trigger an umlaut rule that non-epenthetic *u* does trigger. In the monocluster account, there simply is no vowel until the postcyclic morphology, at which point the *u*-triggered rule no longer applies. While not a telling difference between the present account and Itō's, the stipulation that *r* can head a monocluster word-finally is more straightforward. Final consonantal monoclusters are not uncommon, allowed even at the surface in some languages. We will see an example in Section 7.3, in Cairene Arabic.

7.2.2 *Imdlawn Tashlhiyt Berber*

Before we take up the details of ITB, the interaction of Form Doublet and *Tri must be reexamined. It is commonplace in resyllabification for an unsyllabified vowel to steal the coda of a preceding syllable to provide it with an onset, as we proposed above for Icelandic. Does this happen in iterative syllabification as well? In principle, we can either suppose that Form Doublet is simply blocked by *Tri, as in (35a), or we can suppose that it applies with restructuring (cluster deletion) in order to comply with *Tri, as in (35b), provided of course that the new doublet cluster is permitted by the constraint set. In (35b), *i* must be semi-nuclear, not strongly nuclear.

(35)



ITB gives evidence that Form Cluster applies in the restructuring fashion. The best result would be that this is universal and does not have to be parametrized.

Since I know of no evidence to the contrary, I will assume that Form Cluster always applies in the restructuring fashion.

ITB has the three-vowel inventory $\{a, i, u\}$. The only strongly nuclear phoneme is the low vowel. Remarkably, all other phonemes are semi-nuclear. For this structural inventory $x \gg y$ iff x is more sonorous than y. The default syllabification rule is then:

$$\text{(36)}\quad \text{Timing Slot} \;\; ; \;\text{Clustered} \; ; \begin{pmatrix} \gg \\ \text{Left} \end{pmatrix} :: \begin{bmatrix} \text{Form Doublet} \\ \text{Form Singlet} \end{bmatrix} \; ; \; \{\text{CSI}, {}^*\text{Tri}\}$$

ITB modifies (36) in the following ways:

1. The repair rules are augmented by Adjoin Onset, as in Icelandic, and by Adjoin Coda, the mirror image of Adjoin Onset.
2. The directionality component of the rule order is slightly refined by giving preference to non-initial timing slots of equal sonority.
3. A constraint is added to the constraint set which requires word-initial and word-final nuclei to be sonorants.

The iterative syllabification rule is then:

$$\text{(37)}\quad \text{Timing Slot} \;\; ; \;\text{Clustered} \; ; \begin{pmatrix} \gg \\ {}^*\text{Initial} \\ \text{Left} \end{pmatrix} :: \begin{bmatrix} \text{Form Doublet} \\ \text{Form Singlet} \\ \text{Adjoin Coda} \\ \text{Adjoin Onset} \end{bmatrix} \; ; \; \{{}^*\text{Nonsonorant-edge-nuclei}, \text{CSI}, {}^*\text{Triplet}\}$$

To illustrate the workings of the iterative syllabification rule, several of the syllabification patterns given by Dell and Elmedlaoui, tabulated in (38), are derived below. I follow Dell and Elmedlaoui in capitalizing nuclei in the surface form.

(38)

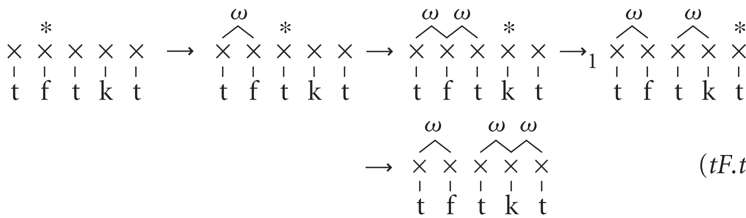
underlying	surface	gloss
txznas	*txZ.nAs*	'store' (3f.sg.perfective, 3m.sg.object)
txznt	*tX.zNt*	'store' (2sg.perfective)
ratlult	*rAt.lUlt*	'you will be born'
tftkt	*tF.tKt*	'you suffered a sprain'
haultn	*hA.uL.tN*	'make them (m.) plentiful'

As a quick reference guide, sonority precedence between the phonemes in the examples which follow is:

$$a > u, i > r, l > m, n > z > f, s, x > b > t, k$$

The first example illustrates non-sonorant nuclei and the restructuring application of Form Doublet.

(39)

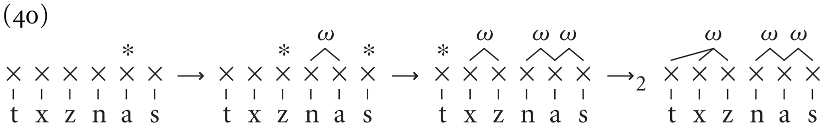


(*tF.tKt*)

In step 1, Form Doublet restructures the preceding syllable.

In the following, sonority increases from left to right, except for the final timing slot.

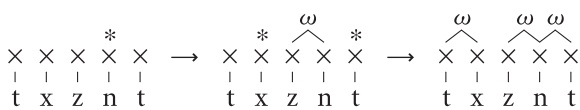
(40)



In step 2, *Nonsonorant-Edge-Nucleus blocks Form Singlet, so Adjoin Onset applies.

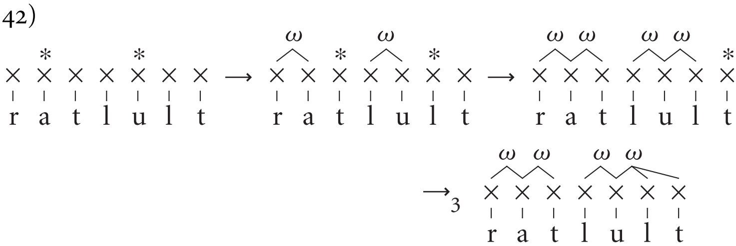
The first four phonemes in (41) are identical to the first four phonemes in (40). Nevertheless, the syllabification turns out to be entirely different.

(41)



We now consider some examples with stretches of falling sonority.

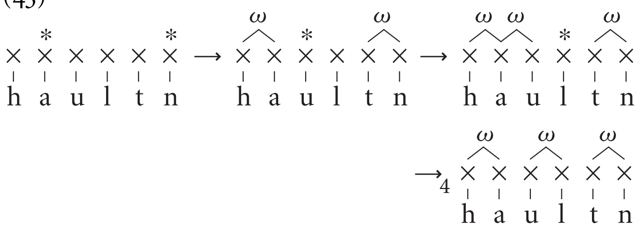
(42)



(*rAt.lUlt*)

In step 3, *Tri blocks Form Doublet and *Nonsonorant-Edge-Nucleus blocks Form Singlet, so Adjoin Coda applies.

(43)

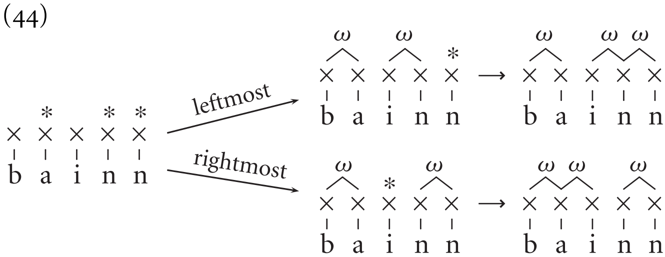


(*hA.wL.tN*)

In step 4, Form Doublet restructures the preceding syllable.

Dell and Elmedlaoui give several examples to show that syllabification works from left to right. Example (44), from their (27), is typical. So that the effect of the secondary determinants of Order can be more easily examined, the local maxima of ≫ (Sonority in ITB) alone are marked, without directionality and without the preference for non-initial timing slots. The example shows that the leftmost of the two adjacent non-initial local maxima of ≫ must be targeted for cluster formation.

(44)

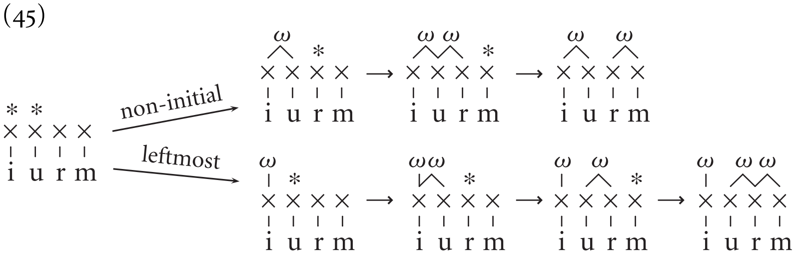


(*bA.yNn*)

(**bAy.nN*)

The following, Dell and Elmedlaoui's (33), shows that *Initial outranks Left in Order. In (45), the first and second timing slots have equal sonority ranking. Choice of the leftmost timing slot incorrectly yields *i.wRm*, while choice of the non-initial timing slot correctly yields *yU.rM*.

(45)



The form *yUy.yL* 'he flew away' in Dell and Elmedlaoui's (42) makes the same point.

7.3 Footing the cluster tier

The view of syllabification developed in the last section leads directly to certain expectations about footing. I assume that footing groups clusters. I assume further that footing not only groups clusters, but must also induce a grouping of the timing slots. This has a major effect on footing. It implies:

(46) *Split-×: Footing delimiters cannot intervene between the two clusters of a bicluster syllable.

If a footing delimiter did intervene, there would be a representation like one of the following:

ω ⟩ ω   ω ⟨ ω
× × ×   or   × × ×

But these are impossible, under the assumption that footing must (indirectly) group timing slots. In both cases, the medial timing slot is both inside a foot and outside it. The constraint (3) is clearly related to Syllable Integrity (Prince 1976). But I do not take *Split-× (or Syllable Integrity) to be a primitive assumption. It is a consequence of the assumption that footing must not only group clusters, but must also induce a grouping of timing slots. The effect of *Split-× on footing is that bicluster syllables on the cluster tier distort the footing pattern that would otherwise obtain (i.e. in the case of all monocluster syllables). A good illustration is the footing pattern in Cairene Arabic.

7.3.1 *Cairene Arabic*

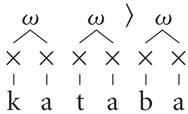
In his analysis of Cairene Arabic, Hayes (1995: 67) points out the "the stress pattern of Arabic as spoken in Cairo has . . . played a central role in the development of metrical theory." I follow that tradition and use Cairene Arabic to illustrate how bicluster syllables affect delimiter insertion. The examples are from Hayes, based on the work of Mitchell (1960), McCarthy (1979), and many others.

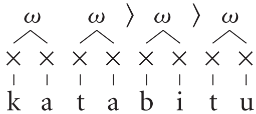
The iterative footing rule (47) for Cairene Arabic is identical to the iterative footing rule given earlier for Southern Paiute, except that "stressable element" is replaced by cluster. As in Southern Paiute, the constraint *Uny is discretionary.

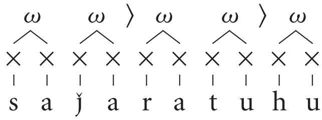
(47) Cluster ; $/\ \omega\_\_ \Rightarrow /\ \_\_\rangle$; Left :: $\emptyset \rightarrow \rangle$; $\{{}^{*}\rangle\% \mid\mid {}^{*}\text{Uny}\}$

*Split-× is taken to be universal, so it need not be mentioned explicitly in (47).

For words consisting of monoclusters, footing proceeds just as in Southern Paiute, producing:

(48) a.  *kátaba* ‘you (m.sg.) wrote’

b.  *katabÌtu* ‘she wrote it (m.)’

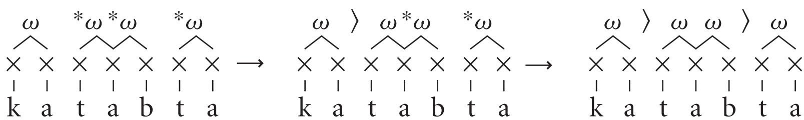
c.  *sağarátuhu* ‘his tree (nom.)’

Although the footing pattern is identical to the footing pattern in similarly structured Southern Paiute words, the stress pattern is different. Southern Paiute has right-foot stress (iambic) and right-word stress, so the main stress is invariably on the penultimate cluster. Cairene Arabic has left-foot stress (trochaic) and right-word stress, so the main stress in words consisting of monoclusters is penultimate if there are an odd number of monoclusters and antepenultimate if there are an even number of monoclusters.

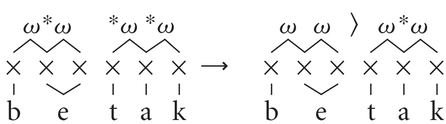
Because *Split-× prevents delimiter insertion between the two clusters of a bicluster syllable (i.e. heavy syllable), heavy syllables in certain positions will distort the pattern of delimiter insertion points. Compare (49a) and (49b) with (48b).

(49)

a. *katábta,* ‘you (m.sg.) wrote’



b. *bé:tak,* ‘your (m.sg.) house’



A violation of discretionary *Uny is forced in (49a). In (49b), there is no way to remove the final defect, which remains. Of course, the defect is forgotten as soon as the iterative footing rule has run its course and the derivation moves on to the next rule.

Now that it is clear how *Split-× affects delimiter insertion, derivations will be written in a more compact manner. The two clusters of biclusters (i.e. heavy syllables) will be visually connected as ω‿ω, indicating that the intercluster position excludes delimiter insertion. Several examples follow in (50). In (50a–d), the position of biclusters is such that *Split-× has no effect. In (50e–g), *Split-× forces violations of *Uny and prevents removal of the right-edge defects.

(50) a. *qat.tá.la,* 'he killed'
ω‿*ω *ω *ω → ω‿ω ⟩ ω *ω → ω‿ω ⟩ ω ⟩ ω

b. *ʔin.ká.sa.ra,* 'it got broken'
ω‿*ω *ω *ω *ω → ω‿ω ⟩ ω *ω *ω → ω‿ω ⟩ ω ω ⟩ ω

c. *ʔad.wi.ya.tú.hu,* 'his drugs (nom.)'
ω‿*ω *ω *ω *ω *ω → ω‿ω ⟩ ω *ω *ω *ω
→ ω‿ω ⟩ ω ω ⟩ ω *ω → ω‿ω ⟩ ω ω ⟩ ω ⟩ ω

d. *haː.ðaː.ni,* 'these (m.du.)'
ω‿*ω *ω‿*ω *ω → ω‿ω ⟩ ω‿*ω *ω → ω‿ω ⟩ ω‿ω ⟩ ω

e. *mu.dár.ris,* 'teacher'
ω *ω‿*ω *ω‿*ω → ω ⟩ ω‿*ω *ω‿*ω → ω ⟩ ω‿ω ⟩ ω‿*ω

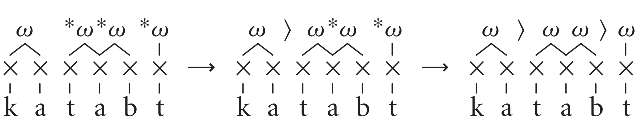
f. *ša.ǰa.rá.tun,* 'tree (nom.)'
ω *ω *ω *ω‿*ω → ω ω ⟩ ω *ω‿*ω → ω ω ⟩ ω ⟩ ω‿*ω

g. *ša.ǰa.ra.tu.hú.maː,* 'their (du.) tree (nom)'
ω *ω *ω *ω *ω *ω‿*ω → ω ω ⟩ ω *ω *ω *ω‿*ω
→ ω ω ⟩ ω ω ⟩ ω *ω‿*ω → ω ω ⟩ ω ω ⟩ ω ⟩ ω‿*ω

Cairene Arabic has what are called "superheavy syllables." What is usually interpreted as a complex coda consonant can appear word at the right edge, as in *katabt*. I assume that Cairene Arabic syllabification does not form complex codas, but does relax the CSI to allow word-final consonantal monoclusters, just as Icelandic allows word-final *r* monoclusters.

We can now derive the feet structure and stress pattern of *katabt*.

(51)



katábt, 'I wrote'

The final monocluster allows the insertion of ⟩ to the right of the bicluster. The result is that if CVC+C is final, stress will always be on the initial cluster of the CVC bicluster.

Aside from needing to clarify the syllable structure for "superheavy syllables," the results above follow virtually without comment. There is no notation of "foot inventory," as in Hayes. Feet are simply what the footing rule produces. There is no notion of extrametricality. Nothing special needed to be said about heavy syllables. Their effect on footing follows entirely from general principles (*Split-×).

7.3.2 *Insensitivity to weight*

It is well known that there are many languages in which footing ignores the distinction between monocluster syllables and bicluster syllables. In other languages, CVV bicluster syllables act as if they have two footing units, but CVC bicluster syllables do not. The theory developed above must be extended to account for these facts.

The topic is extensive and will be more adequately discussed elsewhere, but a sketch of the approach can be given here. A clue as to how this should be treated comes from languages like Central Alaskan Yupik and Malayalam, in which the effective *demotion* of certain bicluster syllables to "monocluster status" for the purposes of footing cannot be fixed in advance of footing, but is an operation which interacts with other footing operations.

This suggests that there is an operation which can effectively "lighten" heavy syllables, which sometimes takes part in iterative footing. Essentially, the operation must render the second cluster of certain (or all) biclusters invisible with respect to footing. Such clusters cannot simply be deleted, because they have a role to play in prosodic structure even though they do not enter the footing calculation. The most straightforward way to achieve invisibility is by splitting the cluster tier into two tiers; a *primary cluster tier* and an *extrametrical cluster tier*. Demotion consists of moving a cluster from the primary cluster tier to the extrametrical cluster tier, preserving all associations. Footing then groups elements on the primary cluster tier. One might imagine that syllable footing, as opposed to cluster footing, arises because a syllable tier is constructed over the cluster tier, and footing takes place on the syllable tier. But splitting the cluster tier is much simpler, involving the introduction of no new elements (i.e. syllables) into the representation.

In languages in which syllable weight plays no role, demotion is applied across the board, prior to footing. In some languages, demotion applies only to CVC syllables. In other languages, demotion has complex interactions with other prosodic rules.

7.3.3 *Another way to remove defects: vowel shortening*

Delimiter insertion is the primary device that iterative footing uses to remove defects. But other devices are also used. Since it is a useful bridge to the discussion of syllable splitting in the next section, it will help to consider one of these devices here.

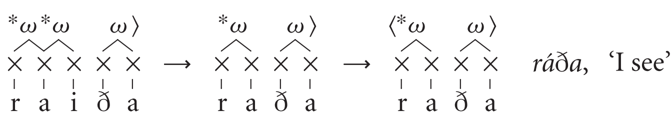
7.3.3.1 *Fijian* Fijian has only (C)V and (C)VV syllables. Footing is carried out by the edge-marking rule (52.EM), followed by the iterative footing rule (52.IF).

(52) EM: $\varnothing \rightarrow \rangle \; / \; \_\_ \%$

$$\text{IF:} \quad \text{Cluster} \;\; ; \;\; \langle\text{-Delimited} \;\; ; \;\; \text{Right} \;\; :: \begin{bmatrix} \varnothing \rightarrow \langle \\ \varnothing \rightarrow \rangle \\ \text{Delete-}\times \end{bmatrix} \; ; \; \{{}^{*}\text{Uny}\}$$

The data is from Hayes (1995: 142), based on Schütz (1985).

Delete-× applies in (53). Neither Ø → ⟩ nor Ø → ⟨ can remove the rightmost defect because of *Split-× and *Uny.

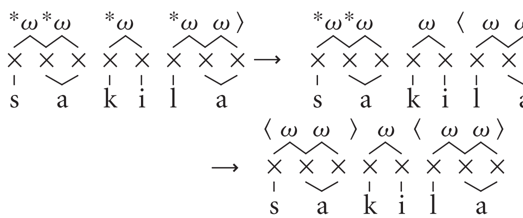
(53) 

Stress is foot-left and main stress is word-right.

Hayes (1995: 145) describes other dialects of Fijian in which the result is *ráiða*, with *ai* a short diphthong. Instead of deletion of the timing slot and its associated phoneme, only the timing slot deletes, with the phoneme reassociating with the timing slot to the left of the deleted slot. In still other dialects, the CVV bicluster splits into a pair of monoclusters. This will be discussed in Section 7.4.

The particular configuration in (53), a bicluster followed by a word-final monocluster, is the only environment in which Delete-× is called on to remove a defect. In all other environments, delimiter insertion suffices. The following example shows why.

(54)



saːkiláː, ‘she knows it’

Crucially, $\varnothing \rightarrow \rangle$ is available along with $\varnothing \rightarrow \langle$, the canonical delimiter insertion rule for right to left footing. The defect-driven rule format allows the iterative rule to alternate between three different ways of removing defects as it progresses across the word. This rule alternation, within the context of a containing iterative rule, distinguishes defect-driven rules from traditional iterative rules, which have no means for such rule alternation.

The distribution of primary and secondary stress, apart from the special configuration in (53), follows with no further comment necessary. Some examples follow, with $^{\mathrm{m}}b$ and $^{\eta}g$ denoting prenasalized consonants.

(55) a. *pa.ràì.ma.ríː* ‘primary’

$$^*\omega \;\; ^*\omega\smile{}^*\omega \;\; ^*\omega \;\; ^*\omega\smile\omega\,\rangle \rightarrow {}^*\omega \;\; ^*\omega\smile{}^*\omega \;\; \omega\,\langle\,\omega\smile\omega\,\rangle$$
$$\rightarrow {}^*\omega \;\; ^*\omega\smile\omega\,\rangle\;\omega\,\langle\,\omega\smile\omega\,\rangle \rightarrow \omega\,\langle\,\omega\smile\omega\,\rangle\;\omega\,\langle\,\omega\smile\omega\,\rangle$$

b. *$^{\mathrm{m}}$bè.le.$^{\mathrm{m}}$bòː.tó.mu*, ‘bell-bottoms’

$$^*\omega \;\; ^*\omega \;\; ^*\omega\smile{}^*\omega \;\; ^*\omega \;\; \omega\,\rangle \rightarrow {}^*\omega \;\; ^*\omega \;\; ^*\omega\smile\omega\,\langle\,\omega \;\; \omega\,\rangle$$
$$\rightarrow {}^*\omega \;\; \omega\,\langle\,\omega\smile\omega\,\langle\,\omega \;\; \omega\,\rangle \rightarrow \langle\,\omega \;\; \omega\,\langle\,\omega\smile\omega\,\langle\,\omega \;\; \omega\,\rangle$$

c. *mìː.sì.ni.$^{\eta}$gá.ni*, ‘machine gun’

$$^*\omega\smile{}^*\omega \;\; ^*\omega \;\; ^*\omega \;\; ^*\omega \;\; \omega\,\rangle \rightarrow {}^*\omega\smile{}^*\omega \;\; ^*\omega \;\; \omega\,\langle\,\omega \;\; \omega\,\rangle$$
$$\rightarrow {}^*\omega\smile\omega\,\langle\,\omega \;\; \omega\,\langle\,\omega \;\; \omega\,\rangle \rightarrow \langle\,\omega\smile\omega\,\langle\,\omega \;\; \omega\,\langle\,\omega \;\; \omega\,\rangle$$

d. *pa.là.si.táː*, ‘plaster’

$$^*\omega \;\; ^*\omega \;\; ^*\omega \;\; ^*\omega\smile\omega\,\rangle \rightarrow {}^*\omega \;\; ^*\omega \;\; \omega\,\langle\,\omega\smile\omega\,\rangle \rightarrow \omega\,\langle\,\omega \;\; \omega\,\langle\,\omega\smile\omega\,\rangle$$

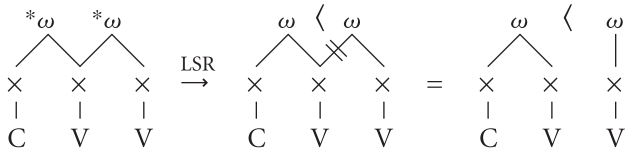
7.4 Syllable splitting

Above, a dialect of Fijian was analyzed in which vowel shortening was used to eliminate the defect (with respect to right to left footing) of the second cluster of a bicluster preceding a word-final monocluster. We begin by analyzing a different dialect of Fijian which eliminates the defect by splitting the bicluster into a pair of monoclusters.

7.4.1 *Syllable splitting in Fijian and Tongan*

*Split-× blocks straightforward delimiter insertion between the two clusters of a bicluster syllable. If delimiter insertion is accompanied by delinking, as in (56), there is no violation of *Split-×. I call the operation *local syllable restructuring* (LSR).

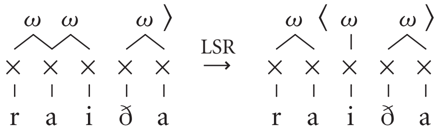
(56)



The operation inserts a delimiter between the two clusters of a bicluster syllable, but avoids a *Split-× violation by simultaneously dissociating the second cluster from the syllable nucleus.

If the crucial example (53) is reconsidered, with LSR replacing Delete-× in the Fijian iterative footing rule (52b), the result is:

(57)



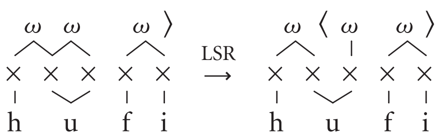
raíða, ‘I see’

This example, Hayes (1995: 145), is from Geraghty (1983).

Tongan is virtually identical to the syllable-splitting dialect of Fijian. Edge marking and iterative footing are as in Fijian, with LSR replacing Delete-× in the iterative footing rule (52.IF). Stress is foot-left and main stress is word-right, as in Fijian. There are differences in secondary stress, which we return to shortly.

Parallel to Fijian (57) is Tongan (58), from Churchward (1953: 11).

(58)



huúfa, ‘to open offcially’

The following examples (Churchward 1953: 5, 11), parallel the Fijian examples in (55), and demonstrate that only biclusters preceding a word-final monocluster are subject to splitting.

(59) a. *kàːká,* ‘to cheat’

$^*\omega\smile{}^*\omega \;\; {}^*\omega\smile\omega\,\rangle \rightarrow {}^*\omega\smile\omega\,\langle\,\omega\smile\omega\,\rangle \rightarrow \langle\,\omega\smile\omega\,\langle\,\omega\smile\omega\,\rangle$

b. *màːlóhi,* ‘strong’

$^*\omega\smile{}^*\omega \;\; {}^*\omega \quad \omega\,\rangle \rightarrow {}^*\omega\smile\omega\,\langle\,\omega \quad \omega\,\rangle \rightarrow \langle\,\omega\smile\omega\,\langle\,\omega \quad \omega\,\rangle$

c. *fàːkahúa,* ‘to sail a zigzag course’

$^*\omega\smile{}^*\omega \;\; {}^*\omega \;\; {}^*\omega\smile\omega \rightarrow {}^*\omega\smile{}^*\omega \;\; {}^*\omega\,\langle\,\omega\smile\omega \rightarrow {}^*\omega\smile\omega\,\rangle\,\omega\,\langle\,\omega\smile\omega$

$\rightarrow \langle\,\omega\smile\omega\,\rangle\,\omega\,\langle\,\omega\smile\omega$

Assuming that foot stress that is not main stress surfaces as secondary stress, the footing rules above predict secondary stress on biclusters and monoclusters which are separated from a following bicluster or a monocluster with main stress by an odd number of monoclusters. Although there are no clear-cut examples given, Churchward (p. 5) indicates that non-penultimate stress only surfaces on long vowels. If so, secondary stress must be suppressed on short vowels. Note however, that the footing rule (52) requires no modification. The only issue is the possible language-particular suppression of certain secondary stresses.

Example (58) was given a prominent place in Prince and Smolensky (1993) as an argument against rule-based phonology.[9] Clearly, in theories which attempt to give an account of how complex phonological representations are computed from simple inputs, syllabification of some form must precede footing. Prince and Smolensky argue that Tongan poses an insoluble “chicken and egg problem” for such theories because the form *huúfi* shows that syllabification depends upon footing. The confusion in the logic of their argument is revealing. The argument is based on the assumption that later operations cannot modify the work of earlier operations. This is an assumption from OT; the idea that all change works “in the same direction”, towards surface optimality. But it is in no way an assumption of rule-based theories. The only substantive point that Prince and Smolensky make is that rule-based analyses of Tongan stress will be unsuccessful if they adopt the premises of Optimality Theory. But there is no argument about this point.

[9] The context was criticism of a proposal of Mester (1992), which had suggested that the Tongan phenomenon was “structure-changing imposition of [a] foot.” My conclusion is that Mester was exactly right. Autosegmental syllabification and an analysis of footing as delimiter insertion into the cluster tier provide an analytic framework which allow Mester’s proposal to be directly translated into a sound analysis.

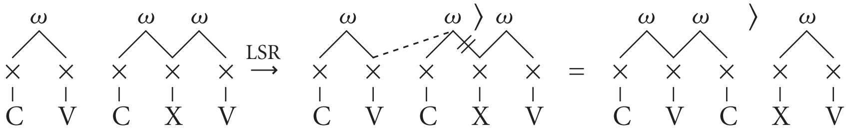
In any event, the LSR rule above has no difficulty in undoing the work that was done in building a CVV bicluster, just as shortening, for example, has no difficulty in undoing the work of associating a long underlying segment with an extra timing slot. That is the way phonology works.

7.4.2 *Gothic (Sievers' Law)*

The metrical structure of Old English and Gothic has been the subject of a series of recent studies. The data below are taken from Dresher and Lahiri (1991), Halle, O'Neil, and Vergnaud (1993), and Keyser and O'Neil (1985).

If a language has melodic segments which can be either onsets or nuclei (X below), it is at least possible for it to employ local syllable restructuring as in (60).

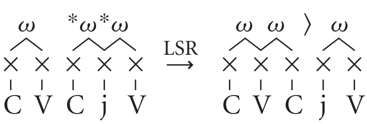
(60)



Formerly nuclear X becomes an onset.

Gothic employs this footing strategy in one very specific context, which is just the left to right version of the context in which vowel shortening and syllable splitting are used in Fijian and Tongan. Just as in Tongan and Fijian, mixed delimiter insertion allows higher-ranked operations to remove footing defects in all other positions, so that resort to LSR can be avoided in those positions. What sets the strategy apart from the Tongan and Fijian footing strategies is that it is employed with only one phoneme, called *j* in the literature, which alternates between an glide onset and high vowel nucleus, depending on the environment it finds itself in. Local syllable restructuring therefore reduces to:

(61)



Nuclear *j* on the left becomes onset *j* on the right.

Prior to iterative footing in Gothic (62.IF), edge marking (62.EM) at the left edge applies.

(62) EM: Ø → ⟨ / # ___

IF: Cluster ; ⟩-Delimited ; Left :: $\begin{bmatrix} \text{Ø} \rightarrow \rangle \\ \text{Ø} \rightarrow \langle \\ \text{LSR} \end{bmatrix}$; {*Uny}

I will assume that syllabification forms complex onsets, but complex codas are forbidden. Final unsyllabified consonants are associated with a monocluster, as in Cairene Arabic. Foot and word stress are left.

First, some examples in which the heavy syllables are distributed in such a way that Ø → ⟩ is sufficient to eliminate all the defects in the foot structure. Subsequent to foot construction, high vowels in unfooted syllables which follow an open syllable delete. Keyser and O'Neil (1985) discovered the connection between high-vowel deletion and foot structure. Their analysis is quite different, but the basic insight is confirmed.

(63)		*underlying*	*foot structure*	*surface*	
	a.	wor.du	⟨ ω‿ω ⟩ ω	*word*	'word'
	b.	ful.wih.tu	⟨ ω‿ω ⟩ ω‿ω ⟩ ω	*fúlwit*	'baptism'
	c.	niː.te.nu	⟨ ω‿ω ⟩ ω ω ⟩	*niːtenu*	'animals'
	d.	oː.þ.er.ne	⟨ ω‿ω ⟩ ω‿ω ⟩ ω	*oːþer*	'other (acc.sg.masc)'
	e.	mi.ki.liis	⟨ ω‿ω ⟩ ω‿ω ⟩	*mikiliis*	'glorify'

Because of *Split-×, biclusters can disrupt footing , as in Cairene Arabic.

(64) a. hea fu dum ⟨ ω‿ω ⟩ ω ⟨ ω‿ω ⟩ *heafdum* 'head'
⟨ ω‿*ω *ω *ω‿*ω → ⟨ ω‿ω ⟩ ω *ω‿*ω → ⟨ ω‿ω ⟩ ω ⟨ ω‿*ω
→ ⟨ ω‿ω ⟩ ω ⟨ ω‿ω ⟩

b. glit mu njis ⟨ ω‿ω ⟩ ω ⟨ ω‿ω ⟩ *glitmuniis*[10] 'glitter'

In (64), biclusters disrupt the footing, but the option Ø → ⟨ allows all defects to be eliminated. But there is one (and only one) configuration in which the option of Ø → ⟨ does not suffice to remove all the defects. If the bicluster immediately follows a word-initial monocluster, and LSR (61) does not apply, *Uny and the edge mark combine to prevent the application of Ø → ⟨.

(65) a. fæ.rel.du 'journey, way'
⟨ ω *ω‿*ω *ω → ⟨ ω *ω‿ω ⟩ ω

[10] The medial high vowel does not delete because the previous syllable is closed. The problem does not appear to have been noted by Halle, O'Neil, and Vergnaud (1993).

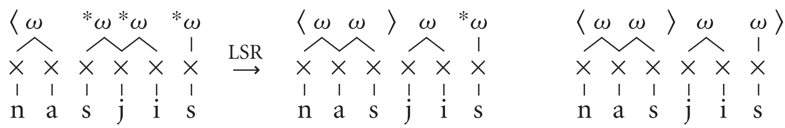
b. si.po:.njis ‘be a disciple’

⟨ ω *ω‿*ω *ω‿*ω → ⟨ ω *ω‿ω ⟩ ω‿*ω → ⟨ ω *ω‿ω ⟩ ω‿ω ⟩

Because *Uny is a strict constraint, one of the defects cannot be removed. This configuration is the mirror image of the configuration which forced syllable splitting in Tongan and Fijian; a bicluster immediately preceding a word-final monocluster.

In the same configuration as (65), if the nucleus of the first cluster in the bicluster is *j*, footing resorts to local syllable reorganization, the bicluster is split, and all defects are removed.

(66)



I leave open the question of whether resyllabification applies in the final foot since there is no evidence that can be brought to bear on the question.

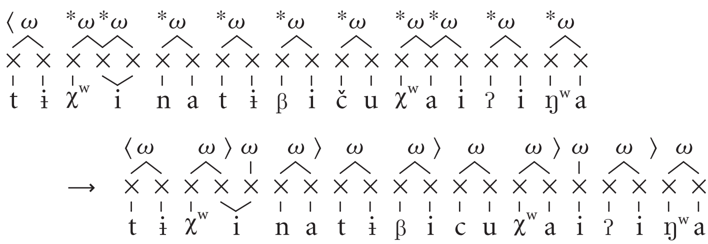
The syllabification of *nasjis* ‘save’ as *nas.jis* rather than *na.siis* is an instance of what is called Sievers’ Law. Another example, with the same analysis, is the syllabification of *arjis* ‘plow’ as *ar.jis*, not *a.riis*. There is a very long tradition of explanation of Sievers’ Law in terms of “exceptional syllabification” of these forms. See the many references in Dresher and Lahiri, starting with Kauffmann (1887). Syllable splitting under accommodation to footing desiderata provides a principled basis for the exceptional syllabification.

7.4.3 *Southern Paiute*

Halle and Vergnaud (1987: 191) note that long vowels never interrupt the metrical count in Southern Paiute, even though long vowels are always counted as a pair of stressable elements. They conclude that, at least as an option, feet can partition syllables. But Hayes (1995: 123), notes that apparent syllable partitioning between two feet is never found with CVC syllables. He further notes that a simple explanation for this is possible if syllables cannot be split between two feet and apparent cases of syllables partitioned between two feet are actually cases of CV.V bisyllabic sequences. Crucially, a C split off from a CVC syllable cannot stand alone as a syllable, so that CV.C is impossible. I find Hayes’s argument convincing, and have adopted *Split-× as a basic principle of foot formation.

In Fijian, Tongan, and Gothic, splitting a bicluster occurred only in a very limited environment. The reason was that both left and right foot delimiters could be inserted by the iterative footing rule. Syllable splitting was always a last resort, and never needed except in a very limited environment. Southern Paiute, which foots left to right, employs only ⟩-insertion. It therefore is faced with the problem of eliminating the footing defects in biclusters throughout the word. Since syllable splitting is available in Southern Paiute and recourse to a less desirable rule is always preferable to violating derivational constraints, even discretionary ones, syllable splitting occurs across the word, when necessary. The following example, Hayes (1995: 121), is from Sapir (1930). The underlying form has two biclusters, but both are broken up by footing. Stress is foot-right and main stress is word-left.

(67)



tɨχʷíːnàtɨβìčuχʷàiʔìŋʷḁ ‘go and ask him to tell a story’.

8

Constraining the Learning Path without Constraints, or The OCP and NoBanana[1]

CHARLES REISS

That which is wanting cannot be numbered.

Ecclesiastes 1: 15

[S]o far as I can tell the story is always more or less the same: whenever there is behavior of significant complexity its most plausible explanation tends to be some explicit process of evolution, not the implicit satisfaction of constraints.

Stephen Wolfram, *A New Kind of Science* (2002: 351)

8.1 Introduction

Many linguists, especially phonologists, have assumed that both Universal Grammar and particular grammars contain constraints, *qua* prohibitions on grammatical structures.[2] However, such prohibitions *cannot* be learned by positive evidence (an infinite number of well-formed structures are absent from the PLD—we may find a supposed ill-formed structure in the next

[1] Thanks to audiences at McGill, UQAM, MIT, Utrecht, and Michigan and to Andrea Gormley, David Odden, Jean-Philippe Marcotte, Daniela Isac, Geoffrey Pullum, Marshall Wong, Brendan Gillon, Patrick Davidson, Bill Idsardi, Madelyn Kissock, Jonathan Bobaljik, Glyne Piggott, Detmar Meurers, Eric Raimy, Ida Toivonen, Ash Asudeh, Morris Halle, and Bert Vaux for useful comments and criticism, not all of which have been addressed. Some of these people remain outraged by the contents of this paper. Many of these ideas were developed in conversations with Mark Hale. It should be clear that many of my proposals draw on various ideas in the literature, some that precede and some that overlap with the seven-year gestation period of this paper. I do not pretend to have approached exhaustive acknowledgement of these sources.

[2] Some developments in Minimalist syntax are discussed below.

sentence we encounter). Therefore, these prohibitions could only be learned *via* negative evidence.

However, it is generally accepted that negative evidence is neither supplied to the child with sufficient regularity, nor attended to by the child enough when supplied, to play a significant role in language learning. Therefore, since the prohibitions cannot be learned *via* positive evidence (for reasons of logic), nor through negative evidence (according to the empirical data), they must be innate.

This conclusion follows from the premises, but I believe it is false. The fault lies with the assumption that UG, and also particular grammars, consist of constraints.

In this chapter, I justify rejection of this premise, and I demonstrate how the need for constraints can be circumvented, while still allowing a learner to converge on a grammar in a finite amount of time. In other words, the learning path is constrained, but not because of language-specific or universal constraints. We thus escape from the tendency to develop overly rich models of Universal Grammar that have culminated in recent theories such as Optimality Theory.

This chapter thus has two goals. One goal is to argue that well-formedness constraints are inappropriate computational devices for modeling grammar. Thus the chapter attempts to do in phonology what recent work by scholars such as Samuel Epstein (Epstein *et al.* 1998; Epstein and Seely forthcoming) is attempting in syntax—to develop a purely derivational theory with minimal theoretical apparatus and no filters or well-formedness constraints. Similar ideas are discussed by Szabolcsi (1988). The conceptual arguments will be bolstered by reference to recent work developing alternative approaches to phonological computation from constraint-based ones.

The second goal is to make concrete proposals concerning the nature of phonological acquisition. The idea is to constrain the acquisition task without recourse to innate constraints.

8.2 The universal NoBanana constraint

Let's turn to a preposterous example. Suppose we are seeking a constrained theory of UG for syntax and we are trying to choose between a theory with the components in (1a) and another with the components in (1b):[3]

[3] I am obviously making simplifying assumptions here. The point is just that one model has a set of entities and the second has all those plus an additional constraint.

(1) Which model of UG is better?

a.
Merge
Lexicon

b.
Merge
Lexicon
NoBanana

Model (a) contains the rule *Merge* which operates on elements of the *Lexicon*. Model (b) contains these components as well as the additional constraint NoBanana which marks as ungrammatical any representation of a sentence containing a banana—an actual banana, not the lexical item *banana*. Is it useful to claim that (b) is a more constrained model than (a) is, since (a) has no way of ruling out sentences that contain bananas? Obviously it is not useful or necessary to do this—(a) does not generate sentences that contain bananas since bananas are not contained in the set of items (the *Lexicon*) over which *Merge* operates. The more constrained model is thus (a), since it is characterized by a subset of the elements needed to characterize (b), and the two models have the same extension.

Consider another preposterous example in (2).

(2) Which model of UG is better?

a.
Merge
Lexicon

b.
Merge
Lexicon
NoLexicalItemsSentToCleveland

In (1) we considered the effect of enriching a model of grammar by adding a constraint referring to entities not found in the set over which *Merge* applies. In (2b), we have added a constraint referring to an operation that is not present in the model of the grammar in (2a). Again, it should be clear that since *Merge* does not have the effect of sending lexical items to Cleveland, and since the grammars characterized in (2) contain no other operations, it is not necessary to rule out representations in which lexical items have been sent to Cleveland.

What makes the preceding examples preposterous is that constraints are supposed to be formulated in terms of a (typically implicit) *universe of discourse*. Note that the claim intended by the constraint NoBanana, that no representation of a sentence contains bananas, is probably true for all human languages. However, there are an infinite number of true claims of this type. No language requires speakers to dance a jig to express iterativity; no language has pizza as an element of syntactic trees; etc. Bananas, pizza, dancing of jigs, sending and Cleveland are not elements of grammatical models. In other words, we do not want our model of grammar to express every true statement about what structures do not occur, since there are an infinite number of such statements and the grammar must be statable in finite terms if it is to be instantiated in human brains.

The conclusion suggested by the preceding discussion is that the search for UG should be conceived of as the attempt to characterize the universe of discourse, the entities and operations that constitute the representations computed by the language faculty. UG is thus to be characterized by a list of categories and rules that take these categories as arguments—and nothing else.

A coherent conception of the "perfection" of the language faculty, one that does not cave in to the temptation of functionalism, is that the formal system that defines UG, as well as every particular grammar, is exhaustively definable: there is a finite list of categories and rules that uniquely determines all and only possible linguistic structures.[4] Again, UG should not be conceived of as a set of constraints defining directly what is *not* a possible human language, because this set has an infinite number of elements. The notion of what is not a possible language will follow from an appropriate characterization of the properties of possible languages, but this notion need not be independently formulated in the grammar.

8.3 Overview

This chapter not only develops this argument concerning what UG should not be, but also makes concrete suggestions concerning how the study of UG should be approached. In Section 8.4, I define constraints in opposition to rules, then I return to the issues raised in the Introduction in order to point out two slightly different ways in which inviolable constraints have been used. I then turn to a discussion of violable constraints, as used in Optimality Theory (Prince and Smolensky 1993). I conclude on philosophical grounds that linguistic theory should be rule-based rather than constraint-based: grammars contain rules (as defined below), not constraints (as defined below).

[4] In other words, the definition of UG, and of particular grammars, can be understood as including a final, exclusion clause of the type used in recursive definitions in logic. I address below the problem of overgeneration—the fact that the set of possible linguistics structures is a superset of attested structures.

In Section 8.5, I briefly show that the ideas presented here converge with some recent work in syntax. I then discuss, in Section 8.6, the use of constraints in conjunction with rule-based phonology, concentrating in Section 8.7 on the Obligatory Contour Principle (Ocp) for illustration. Following Odden (1988) I argue that there is no good theoretical or empirical motivation for positing the Ocp. The argument extends readily to other constraints that have been posited in the literature.

Section 8.8 compares rule- and constraint-based approaches to phonology. I argue for a revival of rule-based phonology, but not a return to the mixing of rules and constraints, and I offer a contribution to the understanding of *formal* aspects of Universal Grammar. The results presented here demonstrate that progress in our understanding of UG does not depend upon the characterization of substantive tendencies subsumed under the notion of markedness. Some ideas concerning the acquisition of phonology, and how learning can be constrained without constraints are presented in section 8.9. Conclusions and open questions are discussed in section 8.10.

8.4 On constraints

This section discusses in general terms various uses of the notion of constraint in linguistic theory. First I discuss the distinction between rules and constraints. Then I discuss constraints on grammars, that is, constraints on what is a possible language. I then turn to inviolable constraints within grammars. Next, I discuss violable constraints as the basis of grammatical computation, as in Optimality Theory. I argue that each of these approaches to defining UG suffers from a combination of a lack of elegance and a mistreatment of the problem of inductive uncertainty.

8.4.1 *What is a rule? What is a constraint?*

Mohanan (2000: 146) argues that, due to basic logical equivalences, the constraint/rule distinction is incoherent once we adopt the view that both rules and constraints express propositions. However, in the following definitions I distinguish rules and constraints both in terms of their role in a computational system (a grammar) as a whole and in terms of their putative "grounding" in phonetics.

8.4.1.1 *A system-internal definition of rules vs. constraints* Various practices in the literature may be at odds with the definitions developed here. This purely terminological issue does not bear on the validity of the dichotomy proposed. So, for example, we may find formal statements that are called "constraints" in the context of a given theoretical framework, but which are in fact examples of what is here called a "rule." In some work, e.g., Karttunen (1993), the terms "(declarative) rules" and "constraints" are used interchangeably.

A RULE *R* can be viewed as a function that maps an input representation *I* defined in terms of a set of representational primitives (features and relations) to an output representation *O* which is defined in terms of the same set of primitives. The application of a rule depends upon a potential input representation matching the structural description of the rule. This representational matching procedure (RMP) outputs two possible results: YES, *I* satisfies the structural description of *R*; or NO, *I* does not satisfy the structural description of *R*. If the output of the RMP is YES, *R* applies and relevant parts of *I* are rewritten as *O*. If the output of the RMP is NO, *I* is not affected.

In a constraint-based theory, constraints also contain RMPs that serve to map an input *I* to one of the two possible results YES or NO, as above. However, for each constraint, one of the two values, YES or NO, maps to a further evaluation called VIOLATION and the other to NOVIOLATION. For example, in various versions of Optimality Theory, syllables that fulfill the condition expressed by "Does the syllable have a coda?" map to YES, and in the case of this constraint, YES maps to the value VIOLATION. Syllables without codas map to NO, which for this constraint maps to NOVIOLATION. For the constraint corresponding to the condition "Does the syllable have an onset?", a syllable with an onset maps to YES, which maps, for this constraint, to NOVIOLATION, whereas an onsetless syllable maps to a NO that, for this constraint, maps to VIOLATION. Perhaps this two-step evaluation of constraints is not necessary, but this characterization does reflect the widespread practice of using both negatively stated ("Don't have a coda") and positively stated ("Have an onset") constraints in the literature.

The use to which this evaluation as VIOLATION or NOVIOLATION is put rests with another part of the computational system. Violation of a constraint is passed on to other parts of the computational system. In theories incorporating inviolable constraints, constraint violation prevents a representation from being evaluated as grammatical. In Optimality Theory the violations are used by EVAL, the evaluation procedure which interprets violation with respect to the relative ranking of the constraints.

To reiterate: a rule is defined as a function from representations to representations; a constraint is defined as a function from representations to the set {VIOLATION, NOVIOLATION}.

The role of constraints in a computational system, telling another part of the system that a representation is somehow ill-formed, is related to the issue of NoBanana, discussed above, as follows. There are an infinite number of ways in which a representation can be ill-formed. We do not want the grammar to have to be able to recognize them all.

8.4.1.2 *The system-external basis of well-formedness constraints* In many constraint-based linguistic theories a crucial aspect of constraint evaluation leading to the equivalent of an output value Violation is the notion of ill-formedness or markedness. This represents the second major problem with constraint-based formalism, as defined here.

Depending on the formulation of a given constraint, either matching or failing to match the structural description of the constraint signals ill-formedness. To use the examples introduced above, a constraint formulated as "Don't have a coda" leads to an evaluation of ill-formedness for a syllable which *has* a coda, but a constraint formulated as "Have an onset" leads to an evaluation of ill-formedness for a syllable which *does not have* an onset. Relative and absolute ill-formedness or markedness evaluations of linguistic representations are ascribed by linguists for grammar-external reasons. Marked or ill-formed structures typically are claimed to have at least one of the following properties:

(3) Markedness criteria

- Relative rarity in the languages of the world
- Late "acquisition" by children (typically referring to the recognizability of a form in child speech)
- Loss in aphasia (typically referring to the recognizability of a form in aphasic speech)
- Relative difficulty of perception (not always experimentally validated)
- Relative difficulty of articulation (again, sometimes based on impressions of what is hard to say)
- Tendency to be lost in language change and to not arise in language change

All of these criteria have been criticized by Hale and Reiss (2000*a*, *b*; see references therein). These works conclude that the best way to gain an understanding of the computational system of phonology is to assume that the phonetic substance (say, the spectral properties of sound waves, or the physiology of articulation) that leads to the construction of phonological entities (say, feature matrices) *never* directly determines how the phonological entities are treated by the computational system. The computational system treats features as arbitrary symbols. What this means is that many of the so-called *phonological universals* (often discussed under the rubric of markedness) are in fact epiphenomena deriving from the interaction of extragrammatical factors like acoustic salience and the nature of language change. Phonology is not and should not be grounded in phonetics since the facts which phonetic grounding is meant to explain can be derived without reference to *phonology*. We return to these issues in Section 8.4.7 and in the critique of the Obligatory Contour Principle later in the chapter.

It should be noted that the papers by Hale and Reiss are by no means unique in arguing against the use of "substance" in determining the limits of grammar. A particularly clear example, pointed out to me by Ash Asudeh, is Kaplan (1995/1987: 346-7):

> A formal theory may have a relatively smooth outline . . . [t]hen you start taking chunks out of it . . . because you claim that no human language or grammar has such and such a property. . . . It's a mistake to carry premature and unjustified substantive hypotheses into our computational and mathematical work, especially if it leads to mathematically complex, even if more restrictive, theories. . . . [W]e should be wary of the seduction of substance.

The issue is related, as well, to discussion in other areas of cognitive science. For example, the almost universally held notion that segments that are (allegedly) complex from the articulatory perspective are also representationally complex or marked seems like a clear parallel to the error discussed by Pylyshyn (2003: 8) concerning work on vision: "the mistake of attributing to a mental representation the properties of what it represents." Again, we can relate our discussion to the issues raised by the proposed NoBanana constraint. Why do bananas make representations bad? Because they are not part of the system under scrutiny. But why should a grammatical constraint refer to something that is not part of grammar at all? More commonplace constraints, such as NoFrontRoundVowel, are as poorly motivated a part of Universal Grammar as NoBanana once we recognize that phonetic substance *cannot* be encoded in the phonology. The *acoustic* properties of front, rounded vowels are not directly accessible to the grammar.

8.4.2 *Karttunen (1993)*

I will not review all the literature debating the status of intermediate levels of representation, distinct from both input and output forms, that has appeared in the history of phonology, especially that focusing on Optimality Theory and its immediate predecessors and contemporaries. However, a few comments concerning the influential article of Karttunen (1993) are in order. I think that careful consideration will show that much of the debate concerning sequential vs. parallel derivation is empty.

Karttunen discusses the fact that phonological rewrite rules can be implemented by a finite state transducer. One advantage of expressing rules in this fashion is that transducers express relations between inputs and outputs bidirectionally, and thus can be more immediately useful in developing processing models for both production and parsing. A further advantage is that transducers corresponding to single rules can be composed into a single transducer that implements a "cascade" of ordered rules (p. 180). Thus, the ordered rule format and the transducer format are alternate means of expressing phonological knowledge. The intermediate representations of traditional rule-based phonology need not have a real-time processing referent—they can be understood as corresponding to the contribution made by each component transducer of a complex transducer.

It is interesting to examine Karttunen's ultimate explanation for a turn to two-level models with neither complex transducers nor ordered rules: "the composition of large rule systems to a single transducer turned out to be unfeasible because of practical limitations. A single transducer encoding the complexities of a language like Finnish was too large for the computers available in the early 1980s" (p. 180). Available computational resources have increased significantly over the last two decades, but in any event, such technological considerations are not obviously relevant to the evaluation of psychological theories, especially to the rejection of models that allow reference to intermediate levels of representation.

Whatever the status of the two-level models that arose from such considerations, there are a few points of interest in the context of this chapter. As Karttunen states, the "most fundamental aspect of the two-level rules is that they are deontic statements about correspondences that are possible, necessary, or prohibited in a certain environment"; they are "modal statements about how a form can, must or must not be realized" (Karttunen 1993: 181). In other words, the rules/constraints of the two-level models that Karttunen discusses are purely formal statements, not grounded in phonetic substance. Thus, the arguments used to motivate such a model do not necessarily extend to markedness-based models like Optimality Theory.

8.4.3 *Constraints on grammars*

It is a commonplace in the linguistic literature to find statements suggesting that a goal of linguistic research is to define UG by formulating the constraints on what is a possible language. This enterprise is typically seen as integral to explaining the paradox of language acquisition, in the following way. If the child is endowed with innate knowledge of the constraints delimiting the set of humanly attainable languages, then the child's hypothesis space is limited. Instead of choosing from the infinite set of (not even necessarily attainable) grammars, the learner need only select from a predetermined subset of those. Of course, we might make this idea more palatable to some by referring to constraints on the learner's ability to make hypotheses, rather than to knowledge of these constraints, but this is just a matter of terminology. I wish to argue that a characterization of UG in terms of such constraints can be at best merely a derivative notion.

It is necessary to stress that I am concerned in this subsection with constraints *on* grammars, not constraints *in* grammars. I am not concerned, for the moment, with evaluating the merits of constraint-based computational systems such as Optimality Theory (Prince and Smolensky 1993) *vis-à-vis* rule-based grammars, for example, although I turn to this topic below.

Instead of the preposterous examples in (1) and (2) above, consider the question of hierarchical structure in syntax. Let's imagine that we want to express the claim that all structure is hierarchically organized as a trait of UG. How should this proposal be formulated? If one seeks to characterize UG by listing constraints on the set of possible languages, then one might say something like "Flat structure is not possible" or "All structure is hierarchical". Again, since UG is instantiated in real brains, it must consist of a finite set of characteristic properties. Note again, that we would actually need an infinite set of constraining statements to characterize UG—those referring to bananas, jigs, etc. Again, there are an infinite number of such constraints on the set of possible languages. In order to avoid having an infinitely long list of constraints, constraint-based theories need a *list* of positive statements of entities (distinctive features, primitive operations like *Merge*, etc.). This list will define the universe of discourse in which we interpret a constraint like "Flat structure is not possible". We see, then, that a theory which formulates linguistic universals in terms of constraints must *also* contain a vocabulary of elements and operations in which those constraints are expressed, or to which they refer. This vocabulary of items and processes is presumably based on empirical observations and inferences. Consider a simpler alternative.

If our current hypothesis concerning UG is stated only in *positive* terms, as statements of what grammars have access to or consist of, without prohibitions or constraints, we can achieve a more economical model. The positive terms are just those entities and operations (features, deletions, insertions, *Merge*, *Move*, etc.) which have been observed empirically or inferred in the course of model construction. When faced with a phenomenon which is not immediately amenable to modeling using existing elements of the vocabulary, scientific methodology (basically Occam's Razor) guides us. We must first try to reduce the new phenomenon to a description in terms of the vocabulary we already have. If this can be shown to be impossible, only then can we justify expanding the vocabulary.

Thus, a "constraining approach" to UG, stated in terms of what is disallowed, requires a set of constraints, as well as a vocabulary which defines the universe of discourse in which the constraints are valid. The alternative proposed here requires only the vocabulary of possible entities and operations, along with the metatheoretic principle of Occam's Razor. The alternative is thus more elegant and should be preferred.

In more concrete terms this means that our theory of UG should consist of the minimum number of primitives that we need to describe the grammars we have seen.[5] Note that we should not be influenced in our search by preconceived notions of simplicity. For example, if we know that we need hierarchical structure for some phenomena, but there exist other phenomena which are ambiguous as to whether they require flat or hierarchical structure, then we should assume that the ambiguous cases also have hierarchical structure. If our current theory of UG contains an operation to generate hierarchical structure from primitive elements, constraints against flat structure will be superfluous. In fact, positive statements like "Structures are organized hierarchically" and "All branching is binary" (assuming they are correct) are also superfluous within the grammar itself, even though they are descriptively accurate, since they are just a reflection of how structure building operations work (see Section 8.5).

The approach advocated here seems to be consistent with that used in science in general. If a physicist observes a "constraint" on the behavior of a particle, say, then s/he posits a set of properties for that particle from which the observed behavior emerges. The constraint thus has the status of a derivative and not primitive aspect of the theory.

8.4.4 *Inviolable constraints in grammars*

It was suggested above that the issues raised thus far are irrelevant to the choice between rule-based and constraint-based computational systems. In a sense this was an overstatement and the discussion above is in fact clearly relevant to a certain class of constraints invoked in versions of Optimality Theory, as well as other models of phonology: constraints that are never violated,

[5] According to Rennison (2000: 138) this principle has, in practice, been more vigorously upheld by proponents of Government Phonology (GP), than by members of other schools of phonology.

either universally or within individual grammars. For the sake of concreteness let's adopt a version of Optimality Theory which assumes that it is never the case that the winning candidate in a derivation, in any language, has crossing association lines.[6] There are several ways to deal with this. One possibility is to claim that there exists a constraint, NoCross, that is part of the OT constraint hierarchy which incurs a mark when a candidate contains crossing association lines. This constraint can be posited to be universally undominated, or rather, universally undominated by a "competing" constraint. A competing constraint which dominated NoCross would be one whose satisfaction could "force" a violation of NoCross in the winning candidate. This possibility can be construed as allowing simplicity in the theory—allow GEN to generate candidates freely, and leave it to universally undominated constraints like NoCross to rule out candidates with no chance of surfacing. However, the simplicity achieved is somewhat illusory.

This approach introduces a complication into the core idea of Optimality Theory, the idea that grammars are defined by constraint hierarchies. If one adopts the view that constraints are universal and innate, then certain constraints, the undominatable ones like NoCross will have to be kept in a separate stratum of the constraint hierarchy, one whose members are not subject to reranking. Equivalently, they can be marked as not susceptible to reranking.

Yet another approach is to claim that these constraints are high-ranked at the initial state of the grammar. According to the claim of Smolensky (1996) and most other scholars, they would therefore start out at the top of the block of initially high-ranked Well-formedness constraints. If one is willing to accept such a scenario,[7] then the undominatable constraints need not be marked as unrerankable, since, by hypothesis, no language ever has evidence that they are dominated. However, the generalization that OT grammars consist of freely rerankable constraints becomes empty, if in fact, some of the constraints are never reranked in any language.

We see then that each of the versions of undominatable constraints proposed here leads to complications in the theory of grammar. An obvious

[6] This is a particularly well-known and easily discussed constraint. However, Local and Coleman (1994) have demonstrated that it is basically contentless.

[7] But see Hale and Reiss (1998) for arguments that it is untenable. They argue that acquisition under such an initial ranking, with Well-formedness constraints outranking Faithfulness constraints, is impossible. They claim that the (normal, rerankable) Well-formedness constraints must start out ranked below the Faithfulness constraints in order to allow the acquisition of a lexicon. If one adopts this assumption, then, the undominated Well-formedness constraints like NoCross would have to be initially ranked in a block separated from all the rerankable Well-formedness constraints, or somehow marked as not rerankable.

alternative is to state the constraints as limitations on GEN. In other words, assume that GEN freely generates—except that it does not generate forms that violate NoCross and other undominatable constraints. But this still fails to solve the need to define the universe of discourse for GEN. We would need constraints on GEN to keep it from generating representations that violate NoCross, but not ones that violate NoBanana, presumably. But GEN has certain properties, it does certain things with inputs, and we should try to characterize those properties. Therefore, it seems preferable to model GEN in such a way that it does not have the capacity to output forms with crossing association lines and other impossible traits (including bananas). In other words, the arguments against constraints on grammars and undominatable or inviolable constraints in grammars are the same—we always need a positive characterization of the formal system we are modeling.

8.4.5 *Free generation and constraints as filters*

The dominatable, or violable, constraints of both standard OT, which assumes universal, innate constraints, and other theories which allow language-specific constraints, do not immediately appear to pose the problems discussed thus far. Such constraints are formal devices for evaluating candidates, but they do not, each on its own, define what is a possible linguistic representation. However, I will argue in this subsection that even a constraint-based grammar which contains violable constraints is to be avoided. In section 8.7, we will see that the original motivation for such constraints may have been empirically and methodologically misguided.

Various theories of grammar, including Optimality Theory and some versions of Minimalism and its predecessors posit a mechanism that allows unconstrained generation of linguistic representations. In OT this device is GEN which, given an input, generates the universal candidate set of possible outputs. In various syntactic theories, an analog to GEN is the "free" concatenation of morphemes, or the "free" application of operations such as *Move α*. A derivation which is thus generated will either satisfy certain conditions at PF and LF, the grammar's interface levels, and thus *converge*; or it will not satisfy those conditions and it will *crash*. Both the OT approach and the free-generation-with-interface-conditions approach in syntax are flawed in the following (related) ways.

First, it is easy to proclaim something like "GEN generates any possible linguistic representation" or "The syntactic component allows *Move α* to apply freely." However, it is not clear what such statements mean. One could argue that the theory of grammar need not be computationally tractable, since grammar models knowledge and does not necessarily map directly to an algorithm for generating grammatical output. However, it does not follow from this that we should immediately aim for a model that we cannot imagine being implemented in the mind. It seems that any implementation of Gen or the syntactic component that incorporates *Move α* will have to be very explicit about what it does. One way to achieve this is to be explicit about what the abstract grammar generates. Second, the *free generation-cum-filters* model stinks somewhat of antimentalism. It basically says "We don't care how the candidate forms are generated, as long as they are generated. One way is as good as the next, as long as they are *extensionally* (empirically) equivalent." This is parallel to the position taken by Quine (1972, discussed by Chomsky 1986) in arguing that it is incoherent to talk about the "correct" grammar among a class of extensionally equivalent ones. In defining I-language, a matter of "individual psychology" as the domain of inquiry for linguistics, Chomsky (1986) argued convincingly that the fact that knowledge of language is instantiated in individual minds/brains means that there is necessarily a "correct" characterization of a speaker's grammar (or grammars). We will see below (Section 8.9.3) that such antimentalism does, in fact, show up in current theorizing. Once one accepts that modules/processes, like Gen and *Move α*, must have a certain set of properties; and that these properties ultimately must be derived from a set of positive statements (a vocabulary); and that these properties can be incorporated into the structural descriptions of rules; it appears to be the case that a procedural, or rule-based approach to grammar that generates a sequence of representations constituting a derivation is to be preferred to a constraint-based, non-derivational theory. In other words, grammars can be understood as complex functions mapping inputs to outputs. A rule-based model just breaks the complex function into simpler components, in order to understand the whole. A theory that incorporates Gen or *Move α* avoids the problem of characterizing the function that is the grammar.

8.4.6 *A New Kind of Linguistics?*

Many of the ideas presented here in support of rule-based grammars appear to be paralleled by claims made in Wolfram's (2002) *A New Kind of Science*. Wolfram has almost nothing to say about cognition, and he certainly did not intend his cellular automata models to apply to human grammars, but the logic is very similar to what I am aiming for and it is thus worthwhile quoting him fairly extensively.

In the following passages, equations can be understood as paralleling systems of constraints, and programs as paralleling systems of rules.

(4) Wolfram (2002: 368) arguing against constraints

> It is in many respects easier to work with programs than with equations. For once one has a program, one can always find out what its behavior will be just by running it. Yet with an equation one may need to do elaborate mathematical analysis in order to find out what behavior it can lead to. It does not help that models based on equations are often stated in a purely implicit form, so that rather than giving an actual procedure for determining how a system will behave–as a program does–they just give constraints on what the behavior must be, and provide no particular guidance about finding out what, if any, behavior will in fact satisfy these constraints.

Wolfram is suggesting that constraint-based analyses may provide a valid level of description, but leave unanswered certain important questions and also leave a certain amount of indeterminacy in understanding the nature of the system the constraints describe.

Basically the same point was made earlier in the book too, but from the perspective of the positive attributes of rule systems:

(5) Wolfram (2002: 342) on benefits of rule systems

> One feature of programs is that they immediately provide explicit rules that can be followed to determine how a system will behave. But in traditional science it is common to try to work instead with constraints that are merely supposed implicitly to force certain behavior to occur.
>
> . . . I gave some examples of constraints, and I showed that constraints do exist that can force quite complex behavior to occur. But despite this, my strong suspicion is that of all the examples of complex behavior that we see in nature almost none can in the end best be explained in terms of constraints. The basic reason for this is that to work out what pattern of behavior will satisfy a given constraint usually seems far too difficult for it to be something that happens routinely in nature.
>
> Many types of constraints . . . have the property that given a specific pattern it is fairly easy to check whether the pattern satisfies the constraints. But the crucial point is that this fact by no means implies that it is necessarily easy to go from the constraints to find a pattern that satisfies them.
>
> The situation is quite different from what happens with explicit evolution rules. For if one knows such rules then these rules immediately yield a procedure for working out what behavior will occur. Yet if one only knows constraints then such constraints do not on their own immediately yield any specific procedure for working out what behavior will occur. In principle one could imagine

> looking at every possible pattern, and then picking out the ones that satisfy the constraints.

Given an input to a rule-based phonology, it is typically straightforward to compute the output (as long as the rules are explicit), whereas the problem of finding just which forms best satisfy a constraint system has proven difficult. These passages appear to be particularly relevant to discussions of the computational tractability of Optimality Theory, such as Idsardi (2006). My point in citing Wolfram is not to appeal to authority, but rather to show that the issues faced by linguists may be fruitfully compared to issues in other sciences where it is possible to simulate and model behavior with a greater degree of control.

8.4.7 *The fallacy of imperfection*

> It ain't why, why, why. It just is.
>
> Van Morrison

In phonology at least, it appears that the obstacle to developing such a theory has been an *a priori* belief in the relative well-formedness of abstract representations based on the never formalized notion of markedness. In other words, even the rule-based phonological literature is rife with constraints which are meant to "motivate" the application of rules that repair structure. In syntax, the tradition of appealing to markedness is more subtle, but it has basically been adapted in that the grammar, or perhaps the processor, is characterized with respect to derivations which "crash," as well as with respect to ones that "converge." Consider for comparison the visual system. Given an input, the visual system is assumed to have certain biases, probably manipulable via the little-understood mechanism of *attention*, but no visual input leads to a failure to assign a representation. It is also not clear what it would mean to say that a given representation generated by the visual system was less well-formed, or more marked than another representation. Presumably the visual system generates representations based on the input it is given, and these representations are unique—they are the best and the worst (or rather, neither best nor worst) that the system generates. Outputs are generated which depend on the input and the state of the system processing the inputs—hardly a controversial view. The same holds true of phonological representations—they are not perfect or imperfect, THEY JUST ARE. Since the violable OT constraints are posited on the basis of cross-linguistic typology, data from child speech and the informal intuition of linguists, it is worth evaluating these criteria. I do so here only briefly. Defining markedness based on cross-linguistic *tendencies* of absolute and implicational patterns of attestation (e.g., If a language has voiced stops, it also has voiceless ones) raises many difficult issues, not least of which is "How do we count?" Do we count tokens? E-languages like "English" or "Chinese"? Grammars?[8] Without an explicit theory of what gets counted, generalizations based on intuitive "statistical" patterns are worthless. Furthermore, at least some of the reported statistical tendencies, such as the more common absence of [p] from voiceless stop inventories, in comparison with [t] and [k], are highly reflective of areal biases in the sampling procedure (see Engstrand (1997) and Hale and Reiss (2000*a, b*) for discussion).

Hale and Reiss (1998) have argued in detail that the use of child speech data to determine markedness status is flawed since this data is rendered opaque by the effects of children's performance systems. I will not repeat these arguments here. Linguists' intuitions concerning "better" (unmarked) and "worse" (marked) structures reflect a confusion of levels of analysis, as well as other conceptual problems. A problem addressed in detail by Hale (2000) is that discussion of the evaluation of "output" forms often fails to distinguish between the output of the grammar (a feature-based representation) and, say, the output of the speaker (an acoustic or articulatory event). As demonstrated most clearly by our ability to construct 3D representations based on a black and white pattern on a printed page, there is a vast gap between physical stimuli and outputs and the representations that relate to them. Therefore, even if phonologists had a metric of the complexity or difficulty inherent in interpreting or creating certain physical stimuli or outputs (which they do not), it is apparent that there is no reason to believe that such a scale would translate straightforwardly to a markedness scale for representations. There is no reason to believe, for example, that the representation of the act of pushing a boulder is more difficult or complex or marked than the representation of the act of pushing a feather. Again, recall Pylyshyn's warning about "the mistake of attributing to a mental representation the properties of what it represents."

[8] I am collapsing Chomsky's discussion of a sociopolitical conception of "language," common in everyday parlance, with the E-language conception which he includes among the scientific approaches to the study of language. The E-language approach treats a language as an external artifact, say a text or corpus of texts, rather than as a knowledge state. This collapse is, I believe, justified and consistent with Chomsky's views, since the decision to include various texts or utterances within a single E-language corpus is typically made on the basis of the everyday sociopolitical notion of language—how else can an E-linguist decide that a set of texts constitutes a single corpus, except by appealing to the pretheoretical notion that they are all French or English or Swahili?

8.4.8 *OT constraints as fallible intuitions*

> We should know that one intrinsic characteristic of a heuristic is that it is *fallible*, and that it may be unjustified.
>
> Massimo Piatelli-Palmarini, *Inevitable Illusions* (1994: 22)

The preceding discussion suggests an explanation of why the constraints of OT are violable. These constraints are for the most part derived from so-called "principles of well-formedness" or "markedness" found in other phonological theories. I propose that these "principles" are actually just the heuristic devices that constitute our intuitions as experienced linguists. For example, we may assume that a sequence like [akra] will more likely have a syllable boundary before the stop-liquid cluster than between the two consonants. This is because we seem to believe, rightly or wrongly (it is hard to imagine how to collect the appropriate statistics under the I-language approach) that the majority of languages "maximize onsets" in such cases and leave the first syllable without a coda. However, both syllabifications are found, for example, in the Ancient Greek dialects. Lacking information to the contrary, it may be useful to assume that the more common syllabification is present in a new, unfamiliar language. This will allow the formulation of hypotheses that may then be tested, and the guess will turn out to be correct more often than not, if our intuitions have any basis. However, we must take care not to confuse our intuitions concerning what happens often with the actual nature of the system under study. Based on our experiences and expectations, we apply our intuitions in attempting to solve the problems involved with analyzing data, but there is no reason to expect that these intuitions directly reflect the nature of the actual mental grammar constructed by a learner. The intuition that heavy things fall faster than light things is very useful when someone drops something from a window, but the intuition needs to be transcended to understand the workings of gravity. Heuristics are used by the analyst to make useful guesses about data, and guesses can be wrong. This is why OT constraints need to be violable—they reflect the fallibility of our guesses.

It may be useful to refer to the error under discussion as a confusion of epistemological issues (concerning the nature of our knowledge) with ontological ones (concerning the nature of phonological systems). One explanation for the pervasiveness of such errors may lie with our terminology. A term like *physics* or *phonology* is used in a systematically ambiguous fashion. *Physics* means both "the study of the properties of the physical world, including gravitational attraction, etc." and "the properties of the physical world, including gravitational attraction, etc." When I fall down the stairs, I do so, not because there is a field of study that concerns itself with gravity, but because of the nature of the physical world, because of gravity itself. l would fall down the stairs even if all the physicists and physics books disappeared—I assume people fell down the stairs before Newton. By failing to make this crucial distinction we can be misled into believing that the *tools* (intuitions) we use in phonology *qua* field of study of the nature of sound systems are constitutive of phonology *qua* the nature of sound systems.

I think the use of violable wellformedness or markedness constraints in OT that are based upon putative statistical tendencies has exactly the status of this kind of error. Reiss (2000) discusses another such case in the OT literature.

8.4.9 *Overgeneration*

Pylyshyn (1984: 205ff) describes a box emitting certain recurrent patterns of signals. He then asks what we can conclude about the nature of the computational mechanism inside the box, based on the observed pattern of output. The answer is that we can conclude nothing, since the observed patterns may reflect the nature of what is being computed (in his example, the output is a Morse Code rendering of English text, and the observed regularity is the "i before e, except after c" rule), not the nature of the computer. In Pylyshyn's words "the observed constraint on [the system's] behavior is due not to its intrinsic capability but to what its states represent." (The observed "constraint" on output vanishes if we input German text, instead of English, since German texts will have words that violate the English spelling rule.) Pylyshyn's example suggests that we should expect our models to overgenerate with respect to the corpus of attested data, since this data is "sifted" by language change, for example. The language faculty may be able to perform computations that we have not observed because of the forms that language data just happen to take. The solution to this situation is clear: posit the minimal theoretical apparatus needed to generate attested patterns, and don't worry too much about overgeneration. We must assume, as a matter of scientific practice, that newly encountered phenomena will be amenable to modeling using current theories. We may be proven wrong, this is in the nature of inductive reasoning. When we are proven wrong we change the assumptions. This is not a bad situation—it just reflects the eternal incompleteness of scientific knowledge.

The fact that we predict the computational possibility of unattested forms is not only possible, but highly likely, given the fact that the language faculty is embedded in a complex system of other cognitive and physiological modules with which it interfaces. Consider the following example. Suppose that the rule $\mathcal{R}$ of a formal system combines the primitive categories of the system $\{a, b, c, d, e\}$ into ordered pairs such as $< a, b >$, $< e, c >$, $< b, d >$, etc. Suppose that after collecting a sample of data we notice that all ordered pairs have occurred except for $< a, d >$. If we then supplement our characterization of the formal system by adding a constraint $^*< a, d >$, what have we gained? We have merely built the descriptive generalization into the grammar. Two preferable alternatives come to mind.

The alternative suggested by Pylyshyn's example is to look outside of the formal system itself. In phonology, for example, the shape of phoneme inventories reflects the nature of sound change and physiological constraints on articulation, not just the cognitive capacity of humans. Not only is it misleading and uninsightful to posit constraints on the formal system that do no more than recapitulate observation, but it also discourages us from looking for a real explanation in a domain other than the characterization of the formal system. (see Hale and Reiss 2000*a, b* for discussion.) This approach is adopted by Reiss (2003a) to account for unattested patterns of quantification in phonological rules.[9]

A second alternative to explore is to examine whether $\mathcal{R}$ has been correctly formulated. Many constraint-based linguistic analyses are built by positing a spurious generalization, then adding constraints to the model to account for the cases which do not match the generalization. It seems more elegant to posit our generalizations more carefully. This approach is taken below in our discussion of so-called OCP effects.

Has the preceding dismissal of concerns of overgeneration made the proposals here vacuous? For example, does the position reduce to the following: "posit a rule that generates all the attested data, and assume that unattested data is the result of accidental gaps in the corpus"? Fortunately, the answer is that this is not the position I am advocating, and this is because of a simple claim that is in direct conflict with general practice, at least in the phonology literature. The claim is that rules are formulated in the *least* general form that is compatible with the data.[10] Generality of application results from lack of specification in structural descriptions; lack of generality, that is, restrictiveness of application results from richly specified structural descriptions. In the

[9] A paper by Bakovic (2005) which criticizes several aspects of this discussion came to my attention as this chapter was going to press. Unfortunately, the careful response that Bakovic's paper deserves cannot be undertaken here.

[10] For example, a palatalization rule that applies before the vowels [i, e] in a language with only the vowels [i, e, a, u, o] should be formulated with the conditioning environment as "before [−back, −round, +tense, −low] vowels", and *not* as "before [−back] vowels".

view of acquisition developed in Reiss (1995, 1999) and Hale and Reiss (2003), it is claimed that representations that are more highly specified than necessary for the purposes of generating target output, are a logical necessity in early grammars. Rules are only made more general, that is with less specified structural descriptions, upon exposure to positive evidence. Therefore, a rule of a particular grammar will generate all and only the data whose representations are subsumed by that encountered during the acquisition process.

8.5 A right-minded approach to syntax

The conclusion to be drawn from the discussion above is that it is in fact best to state our theory of UG in terms of a positive list of what can occur—what Wolfram would perhaps call a list of possible components for programs. This approach actually does delimit the set of possible languages as well as a theory that states constraints on possible linguistic structures, because the normal interpretation of a formal system defined by a set of properties (a vocabulary) is that the system is exhaustively defined by those properties. (See Rennison (2000) for an explicit discussion along these lines.) One can add or subtract one of Euclid's Postulates and explore the consequences of such a move, but any set of postulates is assumed to be exhaustive once stated. Similarly, in physics, new elementary particles are posited only when a phenomenon cannot be accounted for by appeal to those currently identified, or when their existence is predicted on other grounds. Since linguistics posits formal models of (indirectly) observable systems, our current theory is open to revision when forced by new discoveries, but Occam's Razor serves as a check on the current version at any particular time. A model characterized by prohibitions in the form of constraints must implicitly be itself constrained by a vocabulary defining the universe of discourse in which the constraints hold. Therefore, such a model contains a certain amount of unnecessary redundancy.

The derivational approach to syntactic relations developed in Epstein, Groat, Kawashima, and Kitahara (1998) adopts a viewpoint consistent with the "rules only" approach to modeling grammar advocated here. These authors claim (pp. 13–14) that their theory has five innovative properties. The first and the last are most clearly relevant to the discussion in this chapter and can be summarized as follows:

(6) Epstein, Groat, Kawashima, and Kitahara (1998)

- The syntactic computational system consists only of syntactic rules. There are no relations (like *Government*) that are not derivable from the nature of the rules;

- There are no filters or constraints (on non-existent levels of representation such as DS and SS), but only lexical items and operations on these items.

These authors are able, for the most part, to do away with independently stipulated constraints on movement such as Greed and Shortest Move and instead build their effects into the nature of the rule/process *Merge* itself. I understand the goal of this model to be to formulate a rule/process *Merge* which applies in such a way that its outputs are well-formed, as long as it is possible to generate a well-formed output from the current input. Perhaps a better way to describe the model is to say that outputs are "formed", or "not formed", and that the notion "well-formed" is undefined—and unnecessary.

In the rest of this chapter, I explore a parallel approach to phonological derivation. First, I provide some background on the use of constraints within primarily rule-based phonologies. Then I demonstrate the insight that can be gained by building the effects of constraints into the statements of the rules themselves.

8.6 Constraints in rule-based phonology

Despite the fact that phonologists tend to characterize current debate concerning OT as a question of "rules vs. constraints", this is misleading (see Archangeli 1997). Many rule-based analyses make use of constraints such as the Obligatory Contour Principle (Ocp). Constraints in otherwise rule-based phonologies serve two main purposes. Either they define certain structures as disfavored or ill-formed, and thus subject to modification by rule; or they are used to block the application of a rule just in case the rules' output would be disfavored or ill-formed. Work by Paradis (1988*a*) and Calabrese (1988) are typical of the use of constraints as diagnostics for repair of certain structures: if a string satisfies the structural description of a constraint, that is, if it violates the constraint, it must be repaired by a rule. The rule-based account of stress systems presented by Halle and Idsardi (1995) appeals to "Avoidance Constraints" (422ff.) which prevent the application of rules in cases where the rules' output would be a "disfavored" structure. The Ocp has been invoked for both of these purposes in a number of papers, most notably McCarthy (1986) and Yip (1988).

Given the problems with markedness theory alluded to above, note that in the absence of a theory of disfavoredness, this approach is circular: the only real evidence for the disfavored status is that the posited rule appears to be blocked; and the posited reason for the blocking is that the resultant structure would be disfavored. Halle and Idsardi point out that certain advantages derive from mixing rules with constraints in the analysis of individual languages. In general, the use of constraints allows us to formulate simpler rules. However, they note that a fully rule-based analysis is in principle always possible—Halle and Vergnaud (1987) is an example they cite:

> In Halle & Vergnaud (1987), the full metrical constituency was constructed, and at the end disfavored configurations [like stress clash] were eliminated by the application of a rule.

I propose that considerations of elegance for a theory of UG take precedence over elegance in the analysis of individual languages, and thus the Halle and Idsardi system, for example, should be adapted in a way that preserves its mathematical explicitness, while doing away with constraints on unattested structures. A possibility which Halle and Idsardi do not consider[11] is to make the structural descriptions of their rules more complex. As these authors point out, some languages do tolerate stress clash and thus their avoidance constraint is specific to those languages which do not tolerate clash. The rewards of allowing for more complex rules are considerable: constraints become unnecessary and the effects of earlier rules need not be undone.

In brief, Halle and Idsardi need the avoidance constraint AVOID(x(to prevent the generation of Line 0 metrical structures such as (x (x x (x x in a language like Garawa that (1) inserts the leftmost left parenthesis on the basis of an Edge-marking rule, and (2) inserts left parentheses iteratively from the right edge after every second syllable. In a word with an even number of syllables, steps (1) and (2) give, e.g., *(watjim(paŋu*. However, in a word with an odd number of syllables the rules outlined above would generate a "disfavored" (x(structure like *(na(řiŋin(muku(njinam(iřa* where the leftmost syllable has a left parenthesis on both its right and its left. The avoidance constraint blocks the insertion of a parenthesis to the left of the second syllable from the left, and the actually generated Line 0 form is *(nařiŋin(muku(njinam(iřa* with a trisyllabic leftmost constituent. Instead of appealing to an avoidance constraint, the so-called Iterative Constituent Construction rule can be specified to insert a left parenthesis only in the environment x x _ x x. By the normal conventions of interpretation, the structural description is not satisfied by the following structure: x (x _ x x. Thus, the stress-clash configuration is not generated.[12] Again, we cannot rule out such complications to rules *a priori*, without

[11] Idsardi (1992), however, does have a useful discussion of rule-, constraint-, and rule-and-constraint-based approaches to stress.

[12] Because it is not relevant to the discussion, I ignore here the further steps in the derivation, those which follow the construction of the Line 0 structure.

considering that the use of the simpler rule requires adding an additional rule to the grammar (in the Halle and Vergnaud formulation) or else enriching grammatical theory by the use of avoidance constraints (in the Halle and Idsardi formulation).[13]

I thus propose that a goal of future phonological research should be to take the idea of rule-based phonology seriously—by avoiding constraints altogether. Such an approach will offer a principled alternative to Optimality Theory and other constraint-based models. In other words, rather than stating simple, but empirically inadequate rules, reinforced by an arsenal of language-particular or universal constraints, we should attempt to understand what kind of rules we actually need if we are to do without any constraints.

Part of the groundwork for this approach was done over ten years ago in a pair of underappreciated papers by David Odden (1986, 1988). Odden demonstrated that the Ocp is demonstrably *not* a universal constraint on either underlying representations or on the workings of the phonological component. Odden also points out that work appealing to the Ocp is unacceptably vague in defining how, for example, identity of representations is computed. These arguments need not be repeated here, since my goal is to reject the use of all constraints on more general grounds.

8.7 The Obligatory Contour Principle

McCarthy (1986) discusses data from several languages in which a vowel which is expected for independent reasons to be deleted, is instead preserved if its deletion would cause identical consonants to be adjacent: Biblical Hebrew /ka:tab-u:/ → [ka:θvu:] but /sa:bab-u:/ → [sa:vavu:] because deletion would bring together the two underlying [b]s (both of which are spirantized by an unrelated process).[14] The "failure" of the deletion rule to apply is dubbed *antigemination* by McCarthy, since the rule is "blocked" if its application would produce a geminate. McCarthy invokes the Obligatory Contour Principle (Ocp) as the constraint which blocks the rule from applying. This

[13] There are, in fact, other plausible rule-based analyses. Morris Halle (pers. comm.) points out that by first building a single binary foot from the *left* edge of the word, then building binary feet iteratively from the right, the third syllable from the left will remain unfooted in words with an odd number of syllables, but not in those with an even number.

Even number of syllables:	x x) (x x (x x
Odd number of syllables:	x x) x (x x (x x

By projecting the leftmost syllable of each foot, the correct Line 1 configuration is generated for all words.

[14] It has been brought to my attention that vowel length in Hebrew is actually difficult to determine. However, this issue is irrelevant to the point under discussion—any example of "antigemination" will do and additional ones are provided below.

phenomenon involves the failure of deletion rules just in cases where the rule would result in a string of identical adjacent consonants.

Yip (1988) provides a very useful summary, elaboration, and discussion of McCarthy's treatment of the Ocp as a blocker of rules. Consider the following argument:

> If a language has a general phonological rule that is blocked just when the output would contain a sequence of identical feature matrices, we can conclude that the Ocp is operating to constrain derivations ... The alternative is an *ad hoc* condition on such rules, as in [7]:
>
> (7) $A \rightarrow \emptyset / B\_\_C$
> Condition: $B \neq C$
>
> Such a condition not only incurs an additional cost (whereas the Ocp is taken to be universal) but also lacks explanatory power, particularly if contexts B and C are necessary only to state the *ad hoc* condition.

In other words, Yip argues that a theory with language-specific rules and a universal Ocp is a better theory than one with language-specific rules that correctly encode where the rule applies, because adding the necessary conditions to the statement of such rules makes them more complex.

Note that the examples that Yip mentions conform to the first (a) of the following three types of conditions on rule application, but Odden (1988) points out that in fact vowel syncope rules are found with all three of the following types of conditioning:

(8) Some conditions on vowel deletion rules (Odden 1988: 462)

a. Delete a vowel unless flanking Cs are identical.

b. Delete a vowel blindly [whatever the flanking Cs are].

c. Delete a vowel only if flanking Cs are identical.

Condition (a) can be restated as "Delete a vowel if flanking Cs are *not* identical." This is the condition described but rejected by Yip in (7) above: $B \neq C$. But note that Odden's type (c) condition would be written as follows:

(9) Odden's condition (c) in the notation Yip rejects: $B = C$

In other words (a) demands non-identity and (c) demands identity of segments in the structural description of a rule. Thus, there is no reason to propose, as McCarthy and Yip do, that rules that conform to condition (a) illustrate a universal principle of markedness—condition (c) is also a possible rule condition. A rule like (8c) *only* applies when it creates Ocp violations—Odden refers to this phenomenon as *antiantigemination*. So a theory of UG must allow for both types. There is thus no good reason to claim that a universal principle, the OCP, *blocks* deletion in the (a) cases, since deletion can also be *required* in cases that lead to apparent OCP violations when a rule with conditions (b) or (c) applies. Stated in McCarthy's terms (although he does not mention such cases), deletion can be blocked (in case (c)) if the rule will *not* generate an OCP violation. This point was clearly made by Odden, though it seems to have been ignored in most of the subsequent literature.[15]

Note that the logic of attributing cases that fit the profile of (a) to a universal principle and ignoring cases that fit (c), is incoherent. Suppose we examine some data concerning a certain phenomenon and find that all cases fall into two categories, *x* or *y*. If we present only cases of *x* and proclaim that we have found that *x* is always true, then our claim is not valid, *no matter how many positive examples of x we adduce*. The existence of (c) cases makes the existence of (a) cases uninteresting on their own. Odden's observations taken together *are* interesting, as we will see below. Simply put, case (c) is a counterexample to the claim that (a) is universal.[16]

8.7.1 *Treating phonological pathology: The OCP as a rule trigger*

The main point of Yip's paper is that the OCP not only *blocks* rule application as in McCarthy's antigemination cases, but also *triggers* it—it may be the case that a rule applies only to an input that violates the OCP. Instead of an argument based on formal simplicity in rule statements, as discussed above, Yip's discussion of the OCP as a rule trigger illustrates particularly well the assumption that the phonology repairs structures that are somehow pathological—ill-formed or marked or disfavored: "The main contribution of the OCP is that it allows us to separate out condition and cure. The OCP is a trigger, a pressure for change" (74).

In Yip's model the "cure" is effected by language-specific rules. In OT models that make use of similar constraints the "cure" emerges from the constraint ranking. Because of the violability of OT constraints, the winning candidate in an OT derivation is typically not fully "cured'—certain marked structures may be present in the output form.[17] One goal of this chapter is to work towards removing the notion of ill-formedness from the generative component of the phonology. There are representations that are generated,

[15] For example, Keer's (1999) recent OT thesis on the OCP, lists Odden's papers in the bibliography, but makes no reference to them in the text, even in sections discussing antigemination.

[16] Providing a principled response to the reader who finds this discussion to constitute an argument for the violable constraints of Optimality Theory is beyond the scope of this chapter, or perhaps even impossible, reducing to a question of faith.

[17] We might refer to this idea as OT's Fallacy of Imperfection. Imperfection, or markedness, seems to be as irrelevant to linguistic theory as the notion of perfection.

or formed, by grammars; there are representations that are not generated—that is, not formed; but there is no reason to believe that anything a grammar actually generates is ill-formed.

Yip provides a range of examples that show how different solutions can be applied to Ocp violations. They include deletion, dissimilation, and assimilation rules (where assimilation represents multiple linking of a single node, and not identical adjacent nodes). One example of repair by deletion comes from Seri (Marlett and Stemberger 1983). This language has a rule that deletes a coda glottal stop in a syllable with a glottal stop in the onset:

(10) *Seri Glottal Stops*

a. ʔa-a:ʔ-sanx → ʔ-a:-sanx 'who was carried'

b. ʔi-ʔ-a:ʔ-kašni → ʔi-ʔ-a:-kašni 'my being bitten'

c. koʔpanšx 'run like him!'

The rule only applies to tautosyllabic glottal stops so the second glottal stop in (10b) is not affected. In general, coda glottal stops can surface, as shown by (10c).

Yip's account of this process is the following:

[We can] assume that the Laryngeal node is absent except for /ʔ/, and the entries for glottalization in [10ab] are thus adjacent and identical and violate the Ocp. This violation triggers a rule that operates in the domain of the syllable, and the language chooses [one of the possibilities for repairing Ocp violations,] deletion of one matrix (either [+constricted] or [Laryngeal]). The actual rule has four parts, as shown in (11):

(11) *Glottal Degemination*
Domain: Syllable
Tier: Laryngeal
Trigger:
Change: Delete second

The environment is not stated, so the rule is unable to operate unless triggered "from the outside". The outside trigger is, of course, the Ocp, a universal principle and thus free of charge.

In another example, Yip proposes that English uses epenthesis to "cure" Ocp violations of adjacent coronal stridents, thus accounting, for example, for the form of the plural morpheme after coronal stridents: *judges, couches, bushes, cases*, etc. In other words, if epenthesis did not apply, the adjacent coronal stridents would constitute an Ocp violation. As Odden (1988) points out, the Ocp is invoked rather opportunistically—note that it appears to be irrelevant to identity of adjacent [+voiced] specifications in words like *bins, rugs, hills,* *cars.* More seriously, Odden points out that there are rules that insert vowels only when doing so will specifically *not* repair an Ocp violation. This is case (d) below. There are also rules that insert vowels regardless of the nature of the flanking consonants—case (e). And of course, there are rules that, like English epenthesis, depend on the total or partial identity of flanking segments—case (f).

(12) More conditions on vowel insertion rules (Odden 1988: 462)

d. Insert a vowel unless flanking Cs are identical.

e. Insert a vowel blindly [whatever the flanking Cs are].

f. Insert a vowel only if flanking Cs are identical.

Parallel to (a), condition (d) can be restated as "Insert a vowel if flanking Cs are *not* identical." Thus there is no reason to see (f) as reflecting the Ocp as a trigger when (d) shows that rules may be triggered if and only if they *fail* to fix Ocp violations. The existence of rules with conditions (c) and (d) make it unlikely that appealing to the Ocp as either a trigger or blocker of rules is a fruitful endeavor.

8.7.2 *The* Identity *and* Non-identity Conditions

More of Odden's data will be presented below. For now, note that it is equally possible for a rule to generate Ocp violations (c) as it is to repair them (f). And it is equally possible for a rule to be "blocked" from generating Ocp violations (a) as to be blocked from fixing them (d).[18] Since the goal of phonological theory should be to define the set of computationally possible human languages, Odden's observations provide an excellent opportunity to study the purely formal nature of linguistic rules. In the following discussion, we will concentrate on syncope rules as a matter of expository convenience. Again, for expository convenience, we will refer to a schematic representation C_1VC_2. Odden's conditions (a) and (c) can be restated in the following:

(13) The Non-identity Condition on syncope rules (Version 1)
Delete a vowel if flanking Cs are *not* identical ($C_1 \neq C_2$).

(14) The Identity Condition on syncope rules (Version 1)
Delete a vowel if flanking Cs are identical ($C_1 = C_2$).

The apparatus of phonological representation must be at least powerful enough to express the Non-identity Condition and the Identity

[18] Of course, (b) also potentially generates Ocp violations, and (e) potentially repairs Ocp violations.

CONDITION. This issue has implications for feature geometry as a model of phonological representation. There is an insightful discussion of the need for Identity Conditions in Archangeli and Pulleyblank (1994: 368–73). These authors point out that "linked structures themselves are simply one type of configuration involving identity" (369). Archangeli and Pulleyblank present the "Identity Predicate," a relation holding between two arguments, which "is important in a wide variety of phonological contexts" (369). In addition to the OCP cases, they cite the case of Tiv where [+round] spreads between vowels, if and only if they agree in height. Arguments against a linked structure analysis of identity conditions include cases where identity holds across a morpheme boundary—since the identical features belong to different lexical items, they cannot be stored as linked.

In Reiss (2003*a*), I formalize the identity and non-identity conditions and offer further arguments for the inadequacy of a "linked structure" analysis of these conditions. I also argue that autosegmental feature geometry cannot express such conditions, and that a sufficiently powerful formalism makes feature geometry unnecessary, and thus not part of phonological theory.

The crux of the argument against autosegmental representation is that non-identity conditions require that two segments be distinct. This cannot be expressed using just feature geometric association lines. For example, imagine a requirement that C_1 and C_2 be different with respect to some arbitrary feature, that is any feature, or any feature out of a predefined subset of all the features. In other words, the two segments must *not be identical*, but it doesn't matter how they differ. In order to express such a NON-IDENTITY CONDITION we can make use of something like the existential quantifier: there exists at least one feature for which C_1 and C_2 have different values.

8.8 Constraints alone vs. Rules and Constraints vs. Rules alone

A reader may have been convinced by this brief sketch to accept the necessity for the additional power granted to the representational component argued for here—the necessity of quantification—without accepting rejection of constraints. The formulation of constraints that can evaluate identity and non-identity would also require the use of quantification. Therefore, constraints on their own, or constraints in conjunction with rules do not vitiate the need for quantificational statements in grammars.

Consider, however, what we gain by adopting a minimalist approach to characterizing the phonological component in terms of rules: we have a rule component which allows the use of quantificational statements; we have no notion of well-formedness or ill-formedness—the phonology maps inputs to outputs. In the following table I compare three approaches to building a phonology, under the assumption that they are all empirically non-distinct, that is, that they can generate the same sets of output. The Just Rules (JR) approach outlined in this chapter is compared to "standard" OT and a generic Rules and Constraints (RC) model.

(15) Comparison of various approaches to phonology

	OT	RC	JR
a. List of Primitive Entities	yes	yes	yes
b. List of Possible Operations/Functions	yes	yes	yes
c. List of Constraints	yes	yes	no
d. Notion of Ill-formedness	yes	yes	no
e. Notion of Repair	no	yes	no
f. Quantifiers in SDs	yes	yes	yes
g. Representational Matching Procedure	yes	yes	yes

A complete formal theory of phonology must specify what it can generate, so it is necessary to define the universe of discourse by listing the entities (a) and operations (b) that the computations have access to. In OT there are no rules, but as discussed above, a fully explicit version of OT will have to provide a finite characterization of what Gen actually does—a list of possible operations on representations is in fact a necessary part of the model. In addition, OT contains other functions, such as Eval, so all three theories contain functions. The three models cannot be distinguished on these grounds.

Obviously, there are constraints (c) in OT and RC models, and there are none in JR. As Yip explains, the use of constraints presupposes a notion of ill-formedness (d), which I have argued is circular at best, and incoherent at worst, as an explanation of phonological alternation. The constraints are posited on the basis of this intuited sense of well-formedness vs. ill-formedness or markedness. This notion does not exist in the JR model, in which a set of rules maps phonological inputs to outputs.

OT does not prescribe a specific repair (e) for individual markedness violations, but conceives of the grammar as finding an optimal solution across all outputs, which emerges from the ranking. In RC, rules are applied to repair ill-formed structures or to block rule application, thus also appealing to markedness theory. Repair is not part of JR theory. In all three theories, quantifiers (f) are necessary to evaluate the SDs of rules or constraints which refer to identity and non-identity. Similarly, all three theories need some kind of Representational Mapping Procedure to determine which representations satisfy the structural description of its rules or constraints.

Recall that we are assuming that we can compare extensionally equivalent grammars. While straightforward theory comparison is difficult, the "rules only" approach appears to be the most elegant. The list of possible operations is stated in positive terms and thus characterizes the universe of discourse with no additional apparatus. There is no notion of markedness, and thus no reason to conceive of rules as repairing representations. The theory requires rules with a sufficiently rich representational apparatus to define their condition of application. However, as exemplified by the discussion of quantification, this apparatus may be needed by any empirically adequate theory.

8.8.1 *Violability and universality in Optimality Theory*

Optimality Theory is a model of grammar which posits universal, violable constraints that are ranked on a language-particular basis. The universality and the violability of OT constraints are not independent. Obviously, different constraints appear to hold in different languages, so if constraints are universal, they must be violable.

One might maintain an OT-type computational system of ranked constraints while denying the universality of constraints. However, if constraints are not universal, then they must be learned for each language. If they are learned, then they could be learned with appropriate structural descriptions that make them surface-true (ignoring the possibly insurmountable problem of opacity for two-level theories like OT). If they are surface-true, then they need not be violable. In other words, if we weaken the claim to universality of OT constraints the rest of the theoretical edifice of OT crumbles as well.

8.8.2 *Structural descriptions are "constraints" on application*

Let's look back to the type of rule discussed by McCarthy to motivate the restriction of rule application by the OCP. Notice that blocking of a rule R can be achieved in one of two ways—either by applying R and undoing its effects if they are "undesirable," or by "looking ahead" to see what the output would be before applying R, and not applying R if the projected output is undesirable. There is, however, a simpler way of avoiding rule outputs that result in ungrammatical surface forms: reformulate the rule as R', so as to apply only when it should. We have said this much already, however, it is important to realize that the structural description of a rule, the representation that determines whether the rule applies via the representational matching procedure discussed in Section 8.3.1, is nothing other than a constraint on application. McCarthy's rule of vowel syncope in Hebrew applies to vowels between consonants, not to any segment that is between any other two segments. The rule applies only under certain metrical conditions, not under others. The condition that the flanking consonants be non-identical, is thus of the same type as the other constraints on application, the other components of the rule's structural description. In other words, there is no motivation in a rule-based grammar that uses an RMP to also have constraints that are not just part of the structural description of rules.

Analogies may again be useful. There is no reason to assume that a law of Newtonian physics, $f = ma$, that refers to entities like *force*, *mass*, and *acceleration* is actually better seen as a relation between variables $x = yz$, which is constrained by a constraint system that rules out any possible instantiation of $x = yz$ other than $f = ma$. Similarly, a rule or law includes a specification of when it is applicable. Writing highly general rules that lack appropriate structural descriptions to restrict sufficiently when the rules actually apply, and then positing constraints that limit the applicability of a rule seems unproductive.

8.8.3 *What is a possible rule?*

Recall that Yip claims that the fact that OCP "effects" are quite common in the languages of the world should motivate us to remove identity and non-identity conditions from structural descriptions. I suggest that this is exactly the wrong conclusion. These types of conditions are among the most crucial things we need to understand if we want to understand how to characterize the class of possible phonological rules. Ironically, such important empirical work by Yip and McCarthy led to the rejection of rule-based phonology in favor of OT, when it should, instead, have led to a deepening of our understanding of the nature of phonological rules. By appealing to constraints we complicate the theory of grammar unnecessarily, since the RMP used in the structural description of rules already provides the computational power that additional constraints were meant to supply. In addition to this complication we also obscure the question "What is a possible rule?"

8.8.4 *What is Universal Grammar?*

A common characterization of the content of a theory of universal grammar presents the goal of UG theorizing to be a search for properties found in all languages. OT in some sense has solved the problem of UG, thus formulated. All constraints are assumed to be present in all languages; however, because some constraints are outranked by conflicting ones, the effects of the former may not be visible in a particular grammar. For example, all grammars have a constraint FAITHSUC demanding input-output faithfulness for the feature

Suction associated wth clicks. However, in English, it is assumed, the markedness constraint NoSuc outranks the faithfulness constraint, so that clicks would not surface, even if they appeared in an English input representation. Thus we see no evidence for FaithSuc by examining English.

Unfortunately, this approach to universalism seriously misconstrues the nature of theorizing about UG since Chomsky's earliest work. The issue is even discussed as early as Lyons (1970):

(16) Lyons (1970) on Chomskyan UG

- "languages make use of the same formal operations" (p. 115).
- "Chomsky believes that there are certain . . . units that are *universal*, not in the sense that they are necessarily present in all languages, but in the somewhat different and perhaps less usual, sense of the term 'universal,' that they can be defined independently of their occurrence in any particular language and can be identified, when they do occur in particular languages, on the basis of their definition within the general theory" (p. 111).
- "Chomsky accepts that any one of his allegedly universal features might be absent, not only 'from the very next language that becomes accessible,' but also from very many quite familiar languages" (p. 114–15).

Another angle on the Chomskyan view recognizes UG, not as a hypothesis, but as a topic of study, the study of the initial state of the language faculty: "In any computational theory, "learning" can consist only of creating novel combinations of primitives already innately available" (Jackendoff 1990: 40; see also Fodor 1976 and Pylyshyn 1973). Therefore, the OT approach to universalism, which attempts to reduce all language variation to constraint ranking follows from an overly simplistic conception of what UG is. By ascribing all constraints to all languages, OT has solved a problem that derives from a misunderstanding of the nature of the enterprise of UG: "How can we define the 'units' that are present in all languages?" This is a different problem from determining the nature of the human language faculty.

8.9 Constraining learning

The preceding sections were devoted to a critique of the use of constraints in linguistic theorizing. One argument depended on the idea that constraints are epiphenomena that must be interpreted in the context of a more or less explicit characterization of the universe of discourse in which they apply.

A second argument depended on a rejection of the notion of markedness or ill-formedness, on which many constraints are based. We now turn to discussion of how the language learner's search space can be constrained or limited, without recourse to positing constraints as a component of Universal Grammar.

Consider a simple case of allophonic patterning such as the distribution of light and dark laterals in Georgian (Robins and Waterson 1952). The light [l] occurs only before front vowels and the dark [ɫ] occurs elsewhere. The language has five vowels [i,e,u,o,a] so we have several options concerning how to formulate the relevant rule. Let's consider two of them. We could either formulate a rule that said "/ɫ/ > /l/ before non-low front vowels" or a rule that said "/ɫ/ > /l/ before front vowels":

(17) Georgian lateral fronting

- Vowels: [i,e,u,o,a]
- /ɫ/ > [l] before i and e

a. $$\begin{bmatrix} +\text{lateral} \\ +\text{son} \\ \vdots \end{bmatrix} > [-\text{back}] \text{ before } \begin{bmatrix} -\text{back} \\ +\textsc{atr} \\ -\text{low} \\ -\text{round} \end{bmatrix}$$

OR

b. $$\begin{bmatrix} +\text{lateral} \\ +\text{son} \\ \vdots \end{bmatrix} > [-\text{back}] \text{ before } \begin{bmatrix} -\text{back} \end{bmatrix}$$

No language-internal evidence would bear on the matter of selecting the correct formulation of the rule since the only front vowels in the language are non-low. In other words the rules are extensionally equivalent. Despite the fact that we tend to teach beginning students that the second rule is better, since it is more concise, I will argue that the first is the better solution.[19] In the course of this discussion, I hope to convince you that you should care which answer is closer to the correct one, and that justifying this claim is not as hopeless an enterprise as it has seemed to be in the past.

8.9.1 *Approaches to phonology*

Suppose that a child, Junior, is acquiring an "English-type" grammar, the output of which includes forms like [k^hæt]. It seems clear that cognitive scientists,

[19] In Reiss 2003*b* I provided arguments based on cross-linguistic patterns against choosing the most concise formulation of a rule. Here I offer a different kind of evidence leading to a similar conclusion.

phonologists in particular, should set as an ultimate goal finding a solution to the first of the following questions (which is the harder and more interesting one), and they should not be satisfied with merely answering the second.

(18) Two kinds of question

- 'What knowledge state underlies Junior's output such that he says [k^hæt]?'
- 'What is the set of possible knowledge states that could lead to Junior saying [k^hæt]?'

The answer to the first question correctly entails a concern with I-language, language conceived of as knowledge, a matter of "individual psychology" (Chomsky 1986). In other words, phonology is computation over symbolic representations by the phonological component of the mind/brain. Let's refer to this approach as the I-phonology approach. The second is merely concerned with defining extensionally equivalent E-languages, that is language conceived of as sets (or potential sets) of utterances, tokens of behavior. This "E-phonology" approach may involve some interesting theorizing on the formal properties of grammars, both humanly attainable ones and others; however, it cannot be adopted as the right approach for phonology as cognitive science.

I will argue that much of the phonological literature, both before and since the advent of OT, has given up on answering questions of the first type. In fact phonologists have turned away from this goal in at least two ways. "E-phonologists" are concerned with the formal issues entailed by the second type of question. It is important to note, however, that like "I-phonologists" they are concerned with mappings between input and output *representations*.

Others have turned further from the goal of I-phonology in their sometimes tacit rejection of the supposition that phonology is only about knowledge and representations. Instead this work is concerned with more superficial,[20] data-fitting theories of speech output as *behavior*. We can thus refer to this as the "B-phonology" school.

We can characterize the concerns of the three-way distinction we now have with these questions:

(19) Three approaches to phonology

- I-phonology: 'Which humanly attainable knowledge state underlies Junior's computation over phonological representations?'

[20] This word is meant in the sense of "observable," not in a necessarily pejorative sense, although I do believe the approach is misguided.

- E-phonology: 'What is the set of formal systems that would output the same representations as Junior's phonology outputs?'[21]
- B-phonology: 'What can we say about the *sounds* Junior makes?'

The "evaluation procedures" discussed in *SPE* and subsequent work were meant to answer questions of the first type, but Anderson's (1985: 327) remarks on the topic are telling: "Early concern for evaluation procedures ... turned out to be something of a dead end." and "[T]he appeal of feature counting went away ... not with a bang, but with a whimper." I will discuss some simple examples which suggest that prospects for answering the first type of question are not as bleak as they have seemed in the past, and that I-phonology is thus a viable enterprise. I thus attempt to revive these issues by redefining the relationship between the study of phonological theory *per se* and phonological acquisition and learnability.

In addition to making positive proposals, I will point out where other models of phonology have strayed from the pursuit of I-phonology. With respect to OT in particular, I will argue here that the notion of *Richness of the Base* has no place in a theory of I-phonology, and that endowing the learner with an innate set of constraints referring to phonetic substance does nothing to solve the paradox of language acquisition.

8.9.2 *Two reasons to look at acquisition*

Given Kiparsky's (1973: 17) observation that "Children learning their native language do not have the interests of linguists at heart" it is necessary that we view phonology from the learner's perspective. Our reward for such attention to the acquisition process will be twofold. First of all, paying attention to acquisition can tell us what we need *not* worry about. For example, the OT literature is rife with claims of OT's superiority at accounting for conspiracies: "One of the principal reasons that rule-based theory has come under attack is that it offers no satisfactory explanation for conspiracies" (Kager 1997: 463). Kiparsky (1973) has shown, however, that generative phonology does not need the notion of conspiracy. Here is my interpretation of Kiparsky's argument.

[21] There is yet another possible subdistinction: some E-phonologists might concern themselves with only humanly attainable formal systems. I will argue in this chapter that it is useful to assume that given the assumed invariance of the language faculty, only one grammar is attainable on exposure to a given set of input data.

(20) The epiphenomenality of conspiracies (based on Kiparsky 1973: 75ff.)

i. A conspiracy is a set of rules that are "functionally related," that is they lead to the same kinds of output configurations such as "all syllables are open."
ii. If a language has such a set of rules, then the rules of the language will tend to be surface-true (transparent).
iii. Non-transparent (opaque) rules are not surface-true.
iv. Rules that are not surface-true are hard for a learner to learn.
v. Things that are hard to learn are more likely *not* to be learned than things which are easy to learn.
vi. Failure to learn aspects of the ambient language constitutes a diachronic change.
vii. Therefore, (E-)languages are more likely to lose opacity than gain opacity.
viii. Therefore, grammars are likely to look like they have conspiracies.

In other words, the existence of conspiracies is an epiphenomenon due to the fact that languages tend to have transparent rules. This in turn is an epiphenomenon derived from the undeniable fact that individual languages must be learned.

Kiparsky's explanation of conspiracies depends on the fact that acquisition can be *unsuccessful*, resulting in so-called language change (Hale 2007). In other words, tendencies such as "conspiracies" are to be explained by reference to diachronic linguistics where the goal is to define possible changes and to explain why certain changes are more or less likely to occur. We now turn to the question of what *successful* acquisition can potentially tell us.

The second benefit of paying attention to acquisition is that it allows us to take seriously the idea expressed in Chomsky (1986: 3) and elsewhere that Universal Grammar (UG) is the Language Acquisition Device (LAD). In other words, the LAD constrains the set of possible languages by determining how the learner assigns analyses to data provided in the environment, the Primary Linguistic Data (PLD). There are several advantages to such an approach. First, we need no "principles" of UG which are not derivable from, or reducible to, the nature of the LAD. Since we obviously need a learning algorithm (the LAD), a theory with just an LAD is *ceteris paribus* better than a theory with an LAD *and* stipulated principles of UG. This approach also obviates the need for an evaluation metric. Learners never compare extensionally equivalent grammars for simplicity or economy, they just construct the one grammar that is determined by the LAD. This means that there is no reason to introduce the terms "simplicity" and "economy" into the theory since they are contentless labels for aspects of the LAD that are not derivable, that is, they are arbitrary. Note that even if the attempt to collapse UG and the LAD is ultimately misguided, this is not a bad kind of mistake to make. Attempting to collapse the two can lead to the discovery that *some* aspects of our current theory of UG are derivable from the nature of the LAD. Using such findings we can formulate a more streamlined version of UG (*qua* set of stipulated properties of the language faculty not derivable from the LAD) even if we cannot reduce its contents completely.

8.9.3 *Too formal*

In this section I argue, in apparent contradiction to the preceding one, that in some ways, phonologists have been *too* formal in their methods. The contradiction is merely apparent, however, and the problem is mostly one of focus. Since, as Chomsky (1986) puts it, generative lingustics is concerned with matters of "individual psychology" the regularities in the output of linguistic systems need to be seen as the result of innate and learned factors. Focusing on purely formal statements concerning potential *in situ* grammars which are extensionally equivalent misses something critical in that it does not force us to discover *the correct grammar* that constitutes knowledge of some language. Some examples will prove helpful.

Kenstowicz and Kisseberth (1979) provide a useful formulation of Kiparsky's Alternation Condition (AC):

(21) The Alternation Condition (AC) as formulated in Kenstowicz and Kisseberth 1979: 215
Each language has an inventory of segments appearing in underlying representations. Call these segments phonemes. The UR of a morpheme may not contain a phoneme /x/ that is always realized phonetically as identical to the realization of some other phoneme /y/.

We need not worry about which, if any, version of the AC is best, or even if the condition is valid in any form—my point here is one of perspective. If we want to equate UG with the LAD, then, instead of proposing the AC as a principle of UG, we should ask "How does the child set up underlying representations? What is the learning algorithm that is used to capture the apparently real patterns manifested by alternations and distribution of sounds?" Kiparsky (1973) pretty much says this in referring to one version of the AC: "a situation which I termed *absolute neutralization* is either impossible or hard to learn, and should therefore in an explanatory theory of phonology be excluded or specified as not highly valued" (65). The explanatory theory Kiparsky refers to is phonological UG. Once we equate UG and the LAD, Kiparsky's stipulated AC becomes unnecessary. If my suggestion is valid, then it is perhaps unfortunate that later work fails to adopt this position, and the AC is treated as a formal *principle* that constrains grammars (including the lexicon), rather than expressing a generalization about how they are constructed.

It is ironic to note that while a fair amount was written on the AC in the pre-OT era, studies of phonological acquisition posited rules of supposed child phonological systems that violated the AC. For example, kids who do not distinguish [ʃ] from [s] because of a purported rule /ʃ/ > [s] that neutralizes the two, are in blatant violation of the AC. If the AC is conceived as a principle of UG it would be unfortunate if it was violated by children's grammars. A coherent theory that takes acquisition into account will provide a learning algorithm that tells us how underlying representations are generated from the *PLD* (in part by denying the existence of "child phonology rules"—see Hale and Reiss 1998). Therefore such a theory does not need the AC.

Can we relate any of this to OT? One oft-touted property of OT is the notion of richness of the base. Given an appropriate constraint ranking a speaker of English could have any one of a number of forms stored for the lexical item that surfaces as [k^hæt]. For example, they could have /k^hæt/, /kæt/, or /k!æt/. If, say, constraints against clicks and constraints demanding that voiceless stops be aspirated word-initially are ranked high, then all these inputs would surface as [k^hæt]. In other words, the surface inventory is not so much a function of the inputs, but more a result of the ranking. This idea, which is supposed to be as applicable to syntax as it is to phonology, is expressed in discussions of richness of the base in Tesar and Smolensky (1998) and Grimshaw (1997).

> The set of possible inputs to the grammars of all languages is the same. The grammatical inventories of languages are defined as the forms appearing in the structural descriptions that emerge from the grammar when it is fed the universal set of all possible inputs. Thus, systematic differences in inventories arise from different constraint rankings, not different inputs. The lexicon of a language is a sample from the inventory of possible inputs; all properties of the lexicon arise indirectly from the grammar, which delimits the inventory from which the lexicon is drawn. There are no morpheme structure constraints on phonological inputs, no lexical parameter that determines whether a language has *pro*.
>
> (Tesar & Smolensky 1998: 252).

We can also see that it is inevitable that light *do* exists in [English], given the constraint rankings.

(Grimshaw 1997: 387).

We must however ask the following: If the inventory is due to the constraint ranking, then what determines the ranking? The answer is obviously that richness of the base expresses exactly the wrong generalization. The inventory present in the ambient language determines the ranking.

Now it is not a problem that OT with richness of the base would allow apparent violations of the AC (by merging all underlying clicks with plain velars for example) since the AC is not part of the theory. However, who, if not phonologists, will be responsible for deciding whether the child has underlying /k^{h}æt/, /kæt/, or /k!æt/? Since we are interested in I-language, we can (and must) ask which is the correct grammar, not just what is the class of extensionally equivalent descriptively adequate grammars. Recall that the two approaches under consideration correspond to the questions we began with in (18). If we believe that our job ends when we can answer the second question, and that the first is not important or perhaps not even coherent, then we will have sided with the anti-mentalism of Quine on the I-/E-language debate (see Chomsky 1986).

We see then that richness of the base, is actually a SYMPTOM of not having an explicit learning algorithm. It represents an abdication of the responsibility of figuring out what the speaker has stored. Of course, one can attempt to provide OT with an explicit learning algorithm, but then richness of the base becomes irrelevant to a characterization of linguistic knowledge. This characterization of the anti-mentalism implicit in many OT analyses is explicit in McCarthy (1999*a*: 6): "with faithfulness bottom-ranked, the choice of input [among three alternatives] doesn't matter, since all map to [the same surface form]. So there is no need to restrict the inputs." McCarthy is confusing the issue of the linguist designing a grammar, *qua* computational system, with the problem of discovering which *mental* grammar the learner acquires.

There is no question of "restricting" the inputs, but rather a question of figuring out which inputs the learner constructs given the observed data. It is something of a perversion of terms to label our hypothesis about what the LAD does a "restriction," when in fact we mean "selection of a uniquely defined choice."

8.9.4 *Innateness and learnability*

In general, I follow Pinker's (1984/96) formulation of orthodox generative views on acquisition, learnability, and innateness.

8.9.4.1 *The Innateness Hypothesis is a misnomer* One of these orthodox ideas is central to our concerns and therefore must be made explicit. This is the view that the Innateness Hypothesis is something of a misnomer. In fact an innate UG is a logical necessity: "In any computational theory, 'learning' can consist only of creating novel combinations of primitives already innately available" (Jackendoff 1990: 40; cf. Pylyshyn 1973: 33; Fodor 1976; Hale and Reiss 2003 for detailed arguments). Basically, the idea is that if learners do not have the representational apparatus needed to represent input in a given domain such as language, then they can never develop that apparatus since they won't be able to recognize what they are supposed to be recognizing! (except through maturation, which is a kind of innateness). What this means for our purposes is that the child *must* have initial access to the universal phonological feature set. This view is inconsistent with much work on phonological acquisition, but it is the only view consistent with the logical necessity of innate representational primitives. It is obviously also tacitly accepted in versions of OT that assume an innate universal constraint set. This view is also consistent with well-known results concerning infants' ability to distinguish all possible phonetic contrasts. Empirical and logical considerations thus force us to endow the learner with the full representational apparatus provided by UG.

8.9.4.2 *The Subset Principle* If one adopts the standard assumption that children do not make use of negative evidence in the course of language acquisition, one is thereby married to some version of the Subset Principle. In other words, the lack of negative evidence necessitates the early formulation of decreasingly restrictive hypotheses concerning the target grammar. So, the essence of the Subset Principle is that the initial hypothesis S_0 concerning the target grammar is maximally constrained. Hypotheses are more constrained when they are more specific; and they are more specified when they are formulated with relatively richer representations. In other words, the logic of the Subset Principle converges with the logic of the necessary innateness of primitives and the experimental evidence. We have three independent arguments for initial full access. We see then, that initial representations must be very rich; and that acquisition is a process of "pruning," rather than "growing" structure. This view is relevant to both the representation of lexical items and the representations of components of the computational system—the rules or constraints. I now consider these two subcases in turn.

8.9.5 *Two modest examples*

We can now examine the implications of such "initial full access" for learnabilty and acquisition, and ultimately for the nature of mature grammars.

I hope to show that these proposals lead to some simple new insights. As we will see, the results are not strictly tied to the constraint/rule debate and relate instead to issues of representation.

8.9.5.1 *Underspecification* The issue of specificity (i.e., the choice between rules (17ab)) is closely tied to the issue of underspecification in lexical items: an underspecified phonological representation potentially subsumes more tokens than a fully specified one. It is worth pointing out immediately that the notion of underspecification in adult grammars loses some of its appeal as soon as one recognizes that early grammars could not possibly be underspecified. In other words, in the theory we propose, achieving underspecification requires a longer, not a shorter, learning path for the child as "pruner" than it would for the child as "grower." We recognize several mechanisms for achieving some form of underspecification, but they are all very data-oriented—that is, they are forced by exposure to positive evidence. These mechanisms include the following:

(22) Possible sources of underspecification

- Alternations (see Inkelas 1996)
- Patterns "supported" by alternations: the alternation in *a*[k^h]*use* / *a*[k]*usation* allows the /k/ of non-alternating *cat* to be stored without aspiration (Hale and Reiss 1999*a*)
- Phonetic underspecification as evidenced by gradient transitions (Keating 1988)
- Phonetic underspecification as evidenced by "big target spaces" (Hale and Reiss 2003)
- Transparent segments

This list leaves little motivation for the child (and no empirical evidence for the linguist) to posit radical underspecification or redundancy rules. It is worth pointing out that this argument converges with recent work by Inkelas (1996) and Yip (1994) both of whom reject more "philosophical" approaches to underspecification.

8.9.5.2 *The Generality Problem* In this section and the following one, we turn to consider the degree of specification found in the components of the computational system. For the sake of expository clarity, I illustrate with simple *SPE*-style rules and feature matrices.

A simple formulation of one component of the *SPE* evaluation metric is provided by Kenstowicz and Kisseberth (1979):

(23) The Conciseness Condition (K&K: 336)
If there is more than one possible grammar that can be constructed for a given body of data, choose the grammar that is most concise in terms of the number of feature specifications.

Obviously, a more concise rule or set of rules will be *potentially* more general, just as a less concise (more richly specified) one will be more restrictive. Of course, the Conciseness Condition is intended to compare grammars that are extensionally equivalent, so one might suspect that no empirical evidence could possible bear on the issue of what the correct grammar is for a body of data. However, by forming a hypothesis concerning how the grammar is constructed by a learner, we end up with a hypothesis about its mature form. In other words, talk of feature counting need not be the "dead end" that phonologists once felt it to be if we try to take the learner's perspective.

Two relevant subcases for considering the generality of rules can be referred to as the GENERALITY problem and the VACUOUS APPLICATION problem. Anyone who has taught introductory phonology will be familiar with relevant examples. I will illustrate the generality problem using a comparison between the Georgian lateral fronting mentioned above and English voicing assimilation of /z/. The vacuous application problem will be described briefly.

8.9.5.3 *Georgian lateral fronting* As we saw above, Georgian has a five vowels system containing [i,e,u,o,a]. The language has two surface laterals which are in complementary distribution. Plain or clear [l] occurs before the front vowels [i,e]. The velarized back [ɫ] occurs elsewhere. Therefore, it is relatively straightforward to set up a rule of the form in (24).

(24) Georgian

- Vowels: [i,e,u,o,a]
- /ɫ/ > [l] before i and e

But if we try to formalize this, how general/concise do we make the rule? Should it be stated to apply before [-back] or before [-back, -low] vowels? How could we possibly decide?—No empirical language-internal evidence can tell us since the language has no [-back] vowels that aren't [-low], so we have to rely on what a principled learning algorithm will tell us. Another relevant question is "Why do we care?" The answer is that we get paid to care—phonologists are supposed to explain the nature and content of phonological knowledge, a matter of "individual psychology."

Assuming that I have convinced the reader to care, let's proceed to a demonstration of what an explicit learning algorithm will lead us to. First of all, where does such a rule "come from"? The answer is that it is generated on the basis of some kind of positive evidence, that is on the basis of tokens of the rule's application. Let's gloss over some difficult details and imagine that the learner somehow comes up with the generalization that "/ɫ/ > [l] before i" and also with the generalization that "/ɫ/ > [l] before e." The final rule which is acquired is just the result of generalizing across these two "subrules." This process is achieved, of course, by finding the representation which subsumes the two cases—for our purposes, the intersections of the triggering environment will suffice. An early (i.e., rich, highly specified, restrictive) representation of the two subrules is given in (25):

(25) "Subrules" of lateral fronting

a. /ɫ/ > [l] before i

$$\begin{bmatrix} +\text{lateral} \\ +\text{son} \\ \vdots \end{bmatrix} > [\text{-back}] \text{ before } \begin{bmatrix} +\text{hi} \\ +\textsc{atr} \\ -\text{back} \\ -\text{low} \\ -\text{round} \end{bmatrix}$$

AND

b. /ɫ/ > [l] before e

$$\begin{bmatrix} +\text{lateral} \\ +\text{son} \\ \vdots \end{bmatrix} > [\text{-back}] \text{ before } \begin{bmatrix} -\text{hi} \\ +\textsc{atr} \\ -\text{back} \\ -\text{low} \\ -\text{round} \end{bmatrix}$$

The only generalization (loss of specificity) driven by the data is the pruning of the features where the two subrules disagree. The result is shown in (26).

(26) /ɫ/ > [l] before i AND e

$$\begin{bmatrix} +\text{lateral} \\ +\text{son} \\ \vdots \end{bmatrix} > [\text{-back}] \text{ before } \begin{bmatrix} +\textsc{atr} \\ -\text{back} \\ -\text{low} \\ -\text{round} \end{bmatrix}$$

The representation of the environment thus denotes a natural class that includes both [i] and [e], but not [æ]. Therefore, from an acquisition viewpoint, there is no reason to believe that the child does generalize beyond the data (by choosing a *less specified* statement of the rule). This prediction is testable (with certain caveats), e.g., by testing Georgian speakers' production of lateral-æ sequences.[22]

In order to fully develop the ideas here it would be necessary to present a theory of how morphological parsing is achieved over forms that the learner initially stores in unanalyzed form. However, for the sake of explicitness, we can provide the preliminary statement in (27).

(27) How general are rules? (Formulation 1)
The correct statement of a rule arrived at by the *LAD* is the *most highly specified* representation that subsumes all positive instances of the rule.

There is obviously one thing missing from this formulation, namely a guarantee that it does not overgenerate with respect to attested data. In brief, (27) must be reformulated with a qualification, as in (28).

(28) How general are rules? (Formulation 2)
The correct statement of a rule arrived at by the *LAD* is the *most highly specified* representation that subsumes all positive instances of the rule, and subsumes no negative instances of the rule.

The positive and negative instances of the rule are the stored forms which the learner ultimately parses morphologically in the process of figuring out a phonology and a lexicon. Note that (28) is not a description of what the learning algorithm does, but rather a characterization of the rules it generates. In other words, the representation of the Georgian fronting rule that contains specification that the trigger is [−low] is more highly specified than the representation which excludes the specification of [−low]. The more specific, that is, more restrictive, rule is the one provided by the LAD. Of course, this contradicts the common practice of finding the most economical rule.

Are we to conclude from this that the rules of a grammar are never stated in a form which entails greater generality than that provided by a *list* of positive tokens? The answer, due to the nature of our algorithm, is clearly "no." The result will depend on what representations are subsumed by the acquired representation of the rule. We turn now to a case where the rule is predicted to be more general than what might be predicted *a priori* from a list of positive tokens.

8.9.5.4 *English "overgeneralization"* A standard argument for the existence of phonological rules formulated in terms of features is based on the intuition

[22] If they front laterals before, say, [ɪ], this is not necessarily a problem, since [i] may stand for a vowel which includes the [ɪ] space (see Hale and Reiss 2003).

that English speakers will extend the rule that devoices /z/ after voiceless obstruents to apply even after voiceless obstruents that don't occur in English, such as [x] or [ϕ].[23] In other words, since speakers cannot have memorized that the [-s] form of the plural marker, underlying /-z/, occurs after these sounds, it must be the case that speakers generate the correct, voiceless form on the basis of a rule stated in terms of distinctive features. Let's assume that this intuition is in fact valid and that English speakers do, in fact, pluralize *Bach* as [baxs]. This result is trivially predicted by the learning algorithm which creates rules via subsumption. In (29–30) I have broken down the problem in a manner that is meant to aid exposition, and not to reflect, for example, stages of development. Leaving aside the sibilants, English has the following voiceless obstruents, all of which devoice a following /z/ to [s]: [p, t, k, f, θ]. For simplicity, consider what happens when the contexts of devoicing after various stops are compared. These stops all agree in being [−son], [−cont], [−voice], etc. They disagree in place features such as [ant], [lab], and [cor]. So the representation that subsumes all the stops that trigger devoicing does not contain these place features, but does contain the features for which the stops agree. Note that certain features that are typically assumed to be irrelevant, such as [−lat] are also specified, since there is no mechanism to remove them.

(29) Collapsing place of articulation in stops.

$$\begin{bmatrix} -\text{son} \\ -\text{cont} \\ -\text{voice} \\ +\text{ant} \\ +\text{lab} \\ -\text{cor} \\ \vdots \\ -\text{lat} \\ \vdots \end{bmatrix} \cap \begin{bmatrix} -\text{son} \\ -\text{cont} \\ -\text{voice} \\ +\text{ant} \\ -\text{lab} \\ +\text{cor} \\ \vdots \\ -\text{lat} \\ \vdots \end{bmatrix} \cap \begin{bmatrix} -\text{son} \\ -\text{cont} \\ -\text{voice} \\ -\text{ant} \\ -\text{lab} \\ -\text{cor} \\ \vdots \\ -\text{lat} \\ \vdots \end{bmatrix} = \begin{bmatrix} -\text{son} \\ -\text{cont} \\ -\text{voice} \\ \vdots \\ -\text{lat} \\ \vdots \end{bmatrix}$$

Place features can be similarly factored out across the fricatives, generating a representation which is [−son, +cont, −voice ...] without place features. Finally, the general rule is found by collapsing cases for both stops and fricatives, that is, by eliminating [+/−cont].

[23] Let's keep things simple and not worry about the plurals of words ending with coronal stridents, like *bushes, glasses, beaches.*

(30) The trigger of the devoicing rule

$$\begin{bmatrix} -\text{son} \\ -\text{voice} \\ \vdots \\ -\text{lat} \\ \vdots \end{bmatrix}$$

Since the resultant representation of the triggering environment is [−son, −voice] but not specified for [cont] or place features, this representation describes a natural class that includes [x] and [ϕ]. That is, the most highly specified representation that subsumes the actually occurring cases, also subsumes the plurals of constructed English nouns with final [x] and [ϕ].

To summarize, the LAD constructs a rule *R* whose representation subsumes the description of all positive examples of the rule and no negative ones. Presented with a string (a representation) *S* which is not identical to any previously encountered string, *R* will be appear to be generalized (of course, it is just "applied") if and only if *S* is subsumed by the representation of *R*. This is of course what it means to have a rule. If rules did not work this way, then phonology would not show the kind of productivity that the wug-test manifests.

8.9.5.5 *The Vacuous Application Problem* In Reiss (2003*b*), I argue that the correct rule statement for a process like coda devoicing in Russian or Polish must contain the specification [+voice] in the structural description. In other words, the correct formulation of the rule is more like (31b) than (31a).

(31) Two candidates for the coda devoicing rule

a. [+cons, -son] → [−voiced] in CODA

b. [+voiced, +cons, −son] → [-voiced] in CODA

This conclusion is based on the assumption that the LAD provides a single interpretive procedure for structural descriptions, and that (31a) is the rule a learner formulates on exposure to patterns of data like that in (32), seen in languages like Turkish (Inkelas 1996; Inkelas and Orgun 1995). Inkelas argues that there is necessarily a three-way contrast in voicing. Some stem-final stops show a t/d alternation (32a), with [t] appearing in codas and [d] appearing in onsets. Inkelas convincingly argues for an underlying segment that has all the features of a coronal stop, but is unspecified for [voiced]. She denotes this feature bundle as /D/. She states that the segment is assigned the value [−voiced] in codas, and [+voiced] elsewhere. Other stem-final stops consistently surface as [t] and thus are posited to be /t/ underlyingly (32b), and others surface as [d] consistently, and are thus posited to be underlying /d/ (32c).

(32) Turkish voicing alternations

a. Alternating: [Øvoiced] (unmarked for [voiced]) /D/
kanat 'wing' *kanatlar* 'wing-plural' *kanadım* 'wing-1sg.poss'

b. Non-alternating voiceless: [−voiced] /t/
sanat 'art' *sanatlar* 'art-plural' *sanatım* 'art-1sg.poss'

c. Non-alternating voiced: [+voiced] /d/
etüd 'etude' *etüdler* 'etude-plural' *etüdüm* 'etude-1sg.poss'

The point is that the two-way contrast presented to a learner of a language like Russian forces a different rule formulation than that forced by the pattern seen in Turkish. The existence of a single LAD constrains, or rather determines, the hypotheses a learner can make.

8.9.6 *Conclusions on Conciseness*

We can conclude that the Conciseness Condition is not a principle of UG. This is good, since it appears not to work (as recognized even by K&K 1979: 338):

> ...it is not conciseness per se that is involved in giving the correct formulation of a phonological rule. Rather, it is a complex and little understood set of considerations commonly referred to as rule naturalness or optimality.

A rule's conciseness or long-windedness turns out to be determined strictly by the data and the LAD, which makes use of basic set-theoretic operations. The logic of the Subset Principle requires that learners posit the most highly specified rule that is consistent with the data.

By taking seriously the idea that UG is just the LAD we no longer need to appeal to vague notions of markedness, naturalness or optimality. We just have to figure out what formal operations the LAD performs on the representations provided by the transduced data (the PLD).

Finally, it is worth reiterating that certain aspects of phonological knowledge cannot be determined by observing speakers' behavior. However, an explicit theory of the learner's initial state and an explicit theory of the learning algorithm can provide a hypothesis concerning what is not directly observable. Such hypotheses can only be formulated in the context of a theory that takes the competence/performance distinction seriously, that is, a theory of I-phonology.

8.10 Conclusions

It is useful again to make an analogy to see that characterizing UG in terms of constraints on possible grammars, instead of in positive terms, is potentially misguided. When a physicist claims that there are, say, five types of fundamental particles, s/he is not explicitly claiming that no others exist—it is impossible to know everything that exists (inductive uncertainty again). What is being claimed is that all known phenomena (within the relevant domain) can be explained using these five particle types, and so there is no reason to posit any others. Similarly, we can now propose the hypothesis that identity and non-identity conditions can be part of phonological rules, but we do not need to claim that those are the only conditions.

The philosophical arguments against constraints are bolstered by the empirical arguments given in the chapter concerning OCP effects (developed in Reiss 2003*b*). These can be summarized as follows. The invocation of universal constraints depends upon a notion of relative ill-formedness or markedness. Such a notion cannot be justified empirically. There are rules that seem to be blocked if their output would violate the OCP, as well as those that seems to be blocked only if their output would *not* violate the OCP, so there is no reason to grant primacy to one type over the other. So without markedness, universal constraints are unjustified. Language-specific constraints are unnecessary, since their effects can be captured by a more precise formulation of rules.

In Section 8.9, we saw that ridding UG of constraints does not leave the child in a situation where the hypothesis space for language acquisition is unconstrained. By positing maximally restrictive rules, the child can converge on a grammar in a finite amount of time. It is worthwhile to compare the approach proposed here to that presented in an influential pre-OT paper by a phonologist who is one of the most important contributors to the success of OT. McCarthy (1988: 84), in an exposition of feature geometry, states that "The goal of phonology is the construction of a theory in which cross-linguistically common and well-established processes emerge from very simple combinations of the descriptive parameters of the model." For example, "Assimilation is a common process because it is accomplished by an elementary operation of the theory—addition of an association line" (86). After attempting to motivate two operations and two constraints on well-formedness, McCarthy declares that "each operation and constraint is predicted to operate on each class node of the feature geometry in some reasonably well-attested linguistic phenomenon" (90).[24] The vagueness of terms like *common, well-established*, and *reasonably well-attested* should alert us to the lack of rigor inherent in such an approach. A simpler, more explicit approach is to figure out what is the minimum amount of representational and computational machinery needed to generate attested patterns. Rather than seeing this as an original suggestion, it strikes me as "the natural approach: to abstract from the welter of descriptive complexity certain general principles governing computation that would allow the rules of a particular language to be given in very simple forms, with restricted variety" (Chomsky 2000: 122).

With this goal in mind, phonology should not return to the rules-and-constraints models that predate Optimality Theory, but to a pure rule-based formalism. The nature of the types of rules needed by phonological theory thus becomes an empirical question that promises to yield answers if not prejudiced by preconceived notions of what rules "should" look like.

Instead of the taxonomic generalizations offered by spurious markedness-based theories like OT, the approach advocated here will offer deeper insight into the nature of phonological computation. Such insight is the goal of cognitive science in general:

> [I]f we confine ourselves to the scientific and intellectual goals of understanding psychological phenomena [as opposed to predicting observed behavior–cr] one could certainly make a good case for the claim that there is a need to direct our attention away from superficial "data fitting" models toward deeper structural theories (Pylyshyn 1973: 48).

[24] The following sentence is much closer to a coherent proposal: "In other words, we should be able to freely combine the predicates of our theory of representations and our theory of operations and constraints and, in each case, come up with some real rule that languages have." See, however, Hale and Reiss (2000*a*, *b*) for arguments that the set of actually attested languages is expected to be only a subset of the set of computationally possible human languages allowed by UG.

This page intentionally left blank

References

Albright, Adam. 2004. Sub-optimal paradigms in Yiddish. *Proceedings of WCCFL*, 23. Somerville, MA: Cascadilla Press, 1–14.

Alderete, John. 1995. Winnebago accent and Dorsey's Law. In Jill Beckman, Laura Walsh Dickey, and Suzanne Urbanczyk (eds.), *UMOP 18: Papers in Optimality Theory*. Amherst, MA: GLSA, 21–52.

—— 1999*a*. Faithfulness to prosodic heads. ROA 94-000.

—— 1999*b*. Head dependence in stress-epenthesis interaction. In B. Hermans and M. van Oostendorp (eds.), *The Derivational Residue in Phonology*. Amsterdam: John Benjamins, 29–50.

—— 1999*c*. Morphologically governed accent in Optimality Theory. Doctoral dissertation, University of Massachusetts, Amherst MA.

—— Beckman, Jill, Benua, Laura, Gnanadesikan, Amalia, McCarthy, John, and Urbanczyk, Suzanne. 1999. Reduplication with fixed segmentalism. *Linguistic Inquiry*, 30: 327–64.

Altenberg, Evelyn and Vago, Robert. 1983. Theoretical implications of an error analysis of second language phonology production. *Language Learning*, 33: 427–47.

Anderson, Stephen. 1969. West Scandinavian vowel systems and the ordering of phonological rules. Doctoral dissertation, MIT.

—— 1974. *The Organization of Phonology*. New York: Academic Press.

—— 1981. Why phonology isn't "natural". *Linguistic Inquiry* 12: 493–539.

—— 1985. *Phonology in the Twentieth Century: Theories of Rules and Theories of Representations*. Chicago: University of Chicago Press.

—— and Lightfoot, David. 2002. *The Language Organ*. Cambridge: Cambridge University Press.

Anisfeld, Moshe. 1969. Psychological evidence for an intermediate stage in a morphological derivation. *Journal of Verbal Learning and Verbal Behavior*, 8: 191–5.

Antonelli, G. Aldo. 2006. Non-monotonic logic. *Stanford Encyclopedia of Philosophy*.

Anttila, Arto. 1997*a*. Variation in Finnish phonology and morphology. Ph.D. dissertation, Stanford University.

—— 1997*b*. Deriving variation from grammar. In Frans Hinshens, Roeland van Hout, and Leo Wetzels (eds.), *Variation, Change and Phonological Theory*. Amsterdam: John Benjamins, 35–68.

Archangeli, Diana. 1997. Optimality Theory: An introduction to linguistics in the 1990s. In Diana Archangeli and D. Terence Langendoen (eds.), *Optimality Theory: An Overview*. Oxford: Blackwell, 1–32.

—— and Pulleyblank, Douglas. 1994. Grounded Phonology. Cambridge, MA: MIT Press.

Bach, Emmon and Harms, Robert. 1972. How do languages get crazy rules? In R. Stockwell and R. Macawlay (eds.), *Linguistic Change and Generative Theory*. Bloomington: Indiana University Press, 1–21.

Bagemihl, Bruce. 1995. Language games and related areas. In John Goldsmith (ed.), *The Handbook of Phonological Theory*. Oxford: Blackwell, 697–712.

Baković, Eric. 2000. Harmony, dominance, and control. Doctoral dissertation, Rutgers University.

—— 2005. Antigemination, assimilation and the determination of identity. *Phonology*, 22.3: 279–316.

—— 2007. A revised typology of opaque generalizations. ROA 850-0706.

—— and Keer, Edward. 2001. Optionality and ineffability. In Geraldine Legendre, Jane Grimshaw, and Sten Vikner (eds.), *Optimality Theoretic Syntax*. Cambridge, MA: MIT Press. ROA-384-03100.

—— and Wilson, Colin. 2000. Transparency, strict locality, and targeted constraints. *Proceedings of WCCFL*, 19.

Bamberger, Jeanne. 1991. *The mind behind the musical ear: How children develop musical intelligence*. Cambridge, MA: Harvard University Press.

Basri, Hasan. 1997. Phonological nativization of loanwords in Selayarese. Manuscript. SUNY at Stony Brook.

—— 1999. Phonological and Syntactic Reflections of the Morphological Structure of Selayarese. Doctoral dissertation, State University of New York at Stony Brook.

—— Broselow, Ellen, and Finer, Daniel. 1999. Clitics and crisp edges in Makassarese. In Caroline Smallwood and Catherine Kitto (eds.), *AFLA VI: Proceedings of the Sixth International Meeting of the Austronesian Formal Linguistics Association*. Department of Linguistics, University of Toronto, 25–36.

—— —— —— and Selkirk, Elisabeth. 1997*a*. Prosodic word and morphosyntactic word in Makassarese phonology. Talk presented at ZAS Conference on the Prosodic Word.

—— —— —— —— 1997*b*. Phonology-syntax interactions in the Makassar languages. Talk presented at AFLA VII, Vrije Universiteit, Amsterdam.

Benjamin, Geoffrey. 1976. An outline of Temiar grammar. In Philip N. Jenner, Laurence C. Thompson, and Stanley Starosta (eds.), *Austroasiatic Studies*, Part I, 129–87.

Benua, Laura. 1995. Identity effects in morphological truncation. In *University of Massachusetts Occasional Papers in Linguistics, 18: Papers in Optimality Theory*. Amherst: GLSA, University of Massachusetts, 77–136.

—— 1997. Transderivational identity: Phonological relations between words. Ph.D. Dissertation, University of Massachusetts.

Berens, Sally. 1997. The phonetics and phonology of Homshetsma. Bachelor's thesis, Harvard University.

Bergroth, Hugo. 1922. Om kvantitetsförhållandena i den bildade finlandssvenskan. *Studier i nordisk filologi*, 13. Helsingfors.

Bergroth, Hugo. 1928. Finlandssvenska. Handledning till undvikande av provinsialismer i tal och skrift. Second edition. Facsimile reprint 1992 Helsingfors: Schildts. [The first edition appeared in 1917.]

Bermúdez-Otero, Ricardo. 1999. Constraint interaction in language change. Doctoral dissertation, University of Manchester.

—— 2002. Cyclicity or sympathy? A case study. Tenth Manchester Phonology Meeting.

—— 2003. The acquisition of phonological opacity. In Jennifer Spenader, Anders Eriksson, and Östen Dahl (eds.), *Variation within Optimality Theory: Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm: Department of Linguistics, Stockholm University, 25–36.

—— 2006*a*. Phonological domains and opacity effects: a new look at voicing and continuancy in Catalan. Paper given at the workshop 'Approaches to Phonological Opacity', 29th GLOW Colloquium, Barcelona, 5 April 2006. Available at <http://www.bermudez-otero.com/Glow2006.pdf>.

—— 2006*b*. Morphological structure and phonological domains in Spanish denominal derivation. In Sonia Colina and Fernando Martínez-Gil (eds.), *Optimality-Theoretic Studies in Spanish Phonology*. Amsterdam: John Benjamins.

—— 2006*c*. Diachronic phonology. In Paul de Lacy (ed.), *The Cambridge Handbook of Phonology.* Cambridge: Cambridge University Press, 497–517. <http://myweb.tiscali.co.uk/bermudez/handbook.pdf>.

—— 2006*d*. Phonological change in Optimality Theory. In Keith Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn., vol. 9. Oxford: Elsevier, 497–505. <http://myweb.tiscali.co.uk/bermudez/encyclopedia.pdf>.

—— 2007. On the nature of the cycle. Handout from the 15th Manchester Phonology Meeting, Manchester, 25 May. Accessed on 8 June 2007 from <http://myweb.tiscali.co.uk/bermudez/15mfm.pdf>.

—— and Hogg, Richard M. 2003. The actuation problem in Optimality Theory: Phonologisation, rule inversion, and rule loss. In D. Eric Holt (ed.), *Optimality Theory and Language Change.* Dordrecht: Kluwer.

—— and McMahan, April. 2006. English phonology and morphology. In Bas Aarts and April McMahan (eds.), *The Handbook of English Linguistics*. Oxford: Blackwell, 382–410.

Besnard, Philippe., G. Fanselow, and Schaub, Torsten. 2003. Optimality theory as a family of cumulative logics. *Journal of Logic, Language and Information*, 12.2: 153–82.

Bhatia, Tej and Kenstowicz, Michael. 1972. Nasalization in Hindhi: A reconsideration. *Papers in Linguistics*, 5: 202–12.

Blevins, Juliette. 1997. Rules in Optimality Theory. In Iggy Roca (ed.), *Derivations and Constraints in Phonology*. Oxford: Clarendon Press, 227–60.

—— 2004. *Evolutionary Phonology*. Cambridge: Cambridge University Press.

—— and Garrett, Andrew. 2002*a*. Analogical morphophonology. In Kristin Hanson and Sharon Inkelas (eds.), *The Nature of the Word: Essays in honor of Paul Kiparsky*. Cambridge, MA: MIT Press.

Blevins, Juliette 2002*b*. The evolution of metathesis. To appear in a volume edited by Bruce Hayes, Robert Kirchner, and Donca Steriade. Cambridge: Cambridge University Press.

Bliese, L. F. 1981. A generative grammar of Afar. Summer Institute of Linguistics, University of Texas, Arlington.

Bloomfield, Leonard. 1939. Menomini morphophonemics. *Travaux du cercle linguistique de Prague*, 8: 105–15.

Blumenfeld, Lev. 2006. Constraints on phonological interactions. Doctoral dissertation, Stanford University.

Blust, Robert. 2004. Austronesian Nasal Substitution: A Survey. *Oceanic Linguistics*, 43.1: 73–148.

Bobaljik, Jonathan. 1998. Mostly Predictable: Cyclicity and the distribution of schwa in Itelmen. *Proceedings of WECOL*, 9, University of California Santa Cruz. Also available as ROA 208-0797.

—— 2006. Paradigms (optimal and otherwise): A case for scepticism. MIT Paradigms Workshop 2004, revised 2006. Accessed on 2 June 2007 from <http://roa.rutgers.edu/files/856-0806/856-BOBALJIK-0-0.PDF>.

Boersma, Paul. 2000. Learning a grammar in functional phonology. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Phonology, Syntax, and Acquisition in Optimality Theory*. Oxford: Oxford University Press, 465–523.

—— and Hayes, Bruce. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32: 45–86.

Booij, Geert. 1996. Lexical Phonology and the derivational residue. In Jacques Durand and Bernard Laks (eds.), *Current Trends in Phonology: Models and Methods*. European Studies Research Institute and University of Salford, 69–96.

—— 1997. Non-derivational phonology meets lexical phonology. In Iggy Roca (ed.) *Derivations and Constraints in the Lexicon*. Oxford: Clarendon Press, 261–88.

Borowsky, Toni. 1993. On the Word Level. In Sharon Hargus and Ellen Kaisse (eds.), *Studies in Lexical Phonology*. New York: Academic Press, 199–234.

Bradley, Travis. 2006. Spanish Rhotics and Dominican Hypercorrect /s/. *Probus*, 18.1: 1–33.

Brame, Michael. 1970. Arabic phonology: Implications for phonological theory and historical Semitic. Doctoral dissertation, MIT.

Broselow, Ellen. 1982. On the interaction of stress and epenthesis. *Glossa*, 16: 115–32.

—— 1992. Parametric variation in Arabic dialect phonology. In E. Broselow, M. Eid, and J. McCarthy (eds.), *Perspectives on Arabic Linguistics*, IV. Amsterdam & Philadelphia: Johns Benjamins, 7–45.

—— 2000. Stress, epenthesis, and segment transformation in Selayarese loans. *Proceedings of BLS 25*.

—— Chen, Su-I, and Huffman, Marie. 1997. Syllable weight: Convergence of phonology and phonetics. *Phonology*, 14: 47–82.

Buckley, Gene. 1998. Integrity and correspondence in Manam double reduplication. In Pius N. Tamanji and Kiyomi Kusumoto (eds.), *NELS 28: Volume 1, Papers from the Main Sessions*. Amherst: GLSA, University of Massachusetts, 59–68.

Burzio, Luigi. 1998. Multiple correspondences. *Lingua*, 103: 79–109.

Bye, Patrik. 1997. Representing "overlength": Against trimoraic syllables. In Geert Booij and Jeroen van der Weijer (eds.), *Proceedings of the Third HIL Phonology Conference*.

—— 2001. Virtual phonology. Rule sandwiching and multiple opacity in north Saami. Doctoral dissertation, Universitetet i Tromsø.

—— Forthcoming. Allomorphy-selection, not optimalization. In Sylvia Blaho, Patrik Bye, and Martin Krämer (eds.), *Freedom of Analysis*? Berlin: Mouton de Gruyter.

Cable, Seth. 2004. Phonological noun-verb dissimilarities in optimal paradigms. Workshop on Paradigms, MIT 2004. Accessed on 2 June 2007, <http://web.mit.edu/scable/www/work/papers/Phon-N-V-Dissim.pdf>.

Calabrese, Andrea. 1988. Towards a theory of phonological alphabets. Doctoral dissertation, MIT.

—— 1995. A constraint-based theory of phonological markedness and simplification procedures. *Linguistic Inquiry* 26: 373–463.

—— 2005. *Markedness and Economy in a Derivational Model of Phonology*. Berlin: Walter de Gruyter.

Casali, Roderic. 1996. Resolving hiatus. Doctoral dissertation, UCLA.

—— 1997. Vowel elision in hiatus contexts: Which vowel goes? *Language*, 73: 493–533.

Chang, Steve, Plauché, Madelaine C., and Ohala, John. 2001. Markedness and consonant confusion asymmetries. In E. Hume and K. Johnson (eds.), *The Role of Speech Perception in Phonology*. San Diego, CA: Academic Press, 79–101.

Chen-Main, Joan. 2007. Rules, constraints, and overlapping violations: The case of Acoma accent loss. *University of Pennsylvania Working Papers in Linguistics*, 13.1: 29–42.

Chomsky, Noam. 1951. Morphophonemics of Modern Hebrew. Master's thesis, University of Pennsylvania. Published 1979, New York: Garland.

—— 1966. *Topics in the Theory of Generative Grammar*. The Hague: Mouton.

—— 1967. Some general properties of phonological rules. *Language* 43.1: 102–28.

—— 1986. *Knowledge of Language*. Westport, CT: Praeger.

—— 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

—— and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Churchward, Clerk Maxwell. 1953. *Tongan Grammar*. London: Oxford University Press.

Churma, Donald. 1985. *Arguments from External Evidence in Phonology*. New York: Garland.

Clements, G. N. 1990. The role of the sonority cycle in core syllabification. In J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. New York: Cornell University Press, 283–333.

—— 2000. In defense of serialism. *The Linguistic Review* 17.2–4: 81–97.

Coetzee, Andries. 2004. What it means to be a loser: Non-optimal candidates in Optimality Theory. Doctoral dissertation, University of Massachusetts-Amherst.

Collischonn, Gisera. 2007. Can comparative markedness explain counterfeeding opacity in European Portuguese? Paper presented at PAPI 2007, Braga, Portugal.

Coutsougera, Photini. 2000. Opacity and sympathy theory. *Reading Working Papers in Linguistics*, 4: 25–46.

Crowhurst, M. and Hewitt, M. 1998. Conjunctive co-ordination. Manuscript. UNC at Chapel Hill & Brandeis.

Dell, François. 1980. *Generative Phonology and French Phonology*. Cambridge: Cambridge University Press.

—— and Elmedlaoui, Mohamed. 1985. Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *Journal of African Languages and Linguistics*, 7: 105–30.

Dell, Gary, Chang, Franklin, and Griffin, Zenzi. 1999. Connectionist models of language production: lexical access and grammatical encoding. *Cognitive Science*, 23: 517–42.

Diffloth, Gerard. 1976. Expressives in Semai. In Philip N. Jenner, Laurence, C. Thompson, and Stanley Starosta (eds.), *Austroasiatic Studies*, Part I, 249–64.

Dresher, B. Elan. 1996. The rise of Optimality Theory in first-century Palestine. *GLOT International*, 2.1/2: 8.

—— 1999. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, 30: 27–67.

—— and Kaye, Jonathan. 1990. A computational learning model for metrical phonology. *Cognition*, 34: 137–95.

—— and Lahiri, Aditi. 1991. The Germanic foot: Metrical coherence in Old English. *Linguistic Inquiry* 22.2: 251–86.

Dybo, Vladimir. 1977. Zapadnokavkazskaja akcentnaja sistema i problema eë proisxoždenija. In Konferencia Nostratièeskie jazyki i nostratièeskoe jazykoznanie, Tezisy doladov, 41–5.

Eckman, Fred. 1981. On the naturalness of interlanguage phonological rules. *Language Learning*, 31: 195–216.

—— 1984. Universals, typologies and interlanguages. In W. Rutherford (ed.), Language Universals and Second Language Acquisition. Philadelphia: John Benjamins, 79–105.

—— 2005. Typological markedness and second language phonology. Manuscript. University of Wisconsin-Milwaukee.

Edge, Beverley. 1991. The production of word-final voiced obstruents by L1 speakers of Japanese and Cantonese. *Studies in Second Language Acquisition*, 13: 377–93.

Elert, C.-C. 1965. *Phonologic Studies of Quantity in Swedish*. Uppsala: Almquist and Wiksell.

Eliasson, Stig and Lapelle, Nancy. 1973. Generativa regler för svenskans kvantitet. *Arkiv för nordisk filologi*, 88: 133–48.

Ellison, Mark. 2000. The universal constraint set: Convention, not fact. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Phonology, Syntax, and Acquisition in Optimality Theory*. Oxford: Oxford University Press, 524–53.

Engstrand, Olle. 1997. Areal biases in stop paradigms. *Papers from Fonetik 97, The Ninth Swedish Phonetics Conference*, held in Umeå, May 28–30, 1997. Reports from the Department of Phonetics, Umeå University (PHONUM), 4, 187–90.

Epstein, S., Groat, E., Kawashima, R., and Kitahara, H. 1998. *A Derivational Approach to Syntactic Relations*. Oxford: Oxford University Press.

—— and Seely, T. Daniel. Forthcoming. *Transformations and Derivations*. Cambridge: Cambridge University Press.

Erwin, Wallace M. 1963. *A Short Reference Grammar of Iraqi Arabic*. Georgetown: Georgetown University Press.

Fanselow, Gisbert and Féry, Caroline. 2002. Ineffability in grammar. *Linguistische Berichte Sonderheft*, 11: 265–307.

Féry, Caroline and Fanselow, Gisbert. 2002. *A Short Treatise of Optimality Theory. Linguistik in Potsdam*, 18.

Fitzpatrick, Justin. 2006. Sources of Multiple-Reduplication in Salish and beyond. In Shannon T. Bischoff, Lynnika Butler, Peter Norquest, and Daniel Siddiqi (eds.), *Studies in Salishan*, MIT Working Papers on Endangered and Less Familiar Languages, 7, 211–40.

—— and Nevins, Andrew. 2004. Linearization of Nested and Overlapping Precedence in Multiple Reduplication. In Sudha Arunachalam and Tatjana Scheffler (eds.), *Proceedings of the Penn Linguistics Colloquium*, 29, UPenn WPL 10.1: 75–88.

Flemming, Edward. 1995. Auditory features in phonology. Ph.D. Dissertation, UCLA.

—— 2001*a*. Contrast and perceptual distinctiveness. In B. Hayes, R. Kirchner, and D. Steriade (eds.), *The Phonetic Bases of Markedness*. Cambridge: Cambridge University Press.

—— 2001*b*. Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, 18.1: 7–44.

Fodor, Jerry A. 1976. *The Language of Thought*. Sussex: Harvester Press.

Frampton, John. 1999. SPE extensions. Manuscript. Northeastern University.

Fujimura, O., Macchi, M. J., and Lovins, J. B. 1977. Demisyllables and affixes for speech synthesis. Ninth International Congress on Acoustics, Madrid.

Gafos, Diamandis. 1998. A-templatic reduplication. *Linguistic Inquiry*, 29: 515–27.

Gallistel, C. R. 2003. Conditioning from an information processing perspective. *Behavioural Processes*, 61.3: 1234, 1–13.

—— 2006. The nature of learning and the functional architecture of the brain. In Q. Jing, *et al.* (eds.), *Psychological Science Around the World*, vol. 1. Proceedings of the 28th International Congress of Psychology. Sussex: Psychology Press, 63–71.

Geraghty, Paul. 1983. *The History of the Fijian Languages*, Oceanic Linguistics Special Publication, 19. Honolulu: University of Hawaii Press.

Gil, David. 1990. Speaking backwards in Tagalog. Paper presented at 8th ASANAL International Conference, Kuala Lumpur.

Goldrick, Matthew. 2001. Turbid output representations and the unity of opacity. In M. Hirotani, A. Coetzee, N. Hall, and J.-Y. Kim (eds.). *Proceedings of the 30th Annual Meeting of the North East Linguistics Society*, I (Amherst, MA: GLSA), 231–45.

Goldsmith, John. 1989. Autosegmental licensing, inalterability, and harmonic application. CLS 25. Chicago: Chicago Linguistic Society.
—— 1990. *Autosegmental and Metrical Phonology*. Oxford: Blackwell.
—— 1993. Harmonic phonology. In John Goldsmith (ed.), *The Last Phonological Rule: Reflections on Constraints and Derivations*. Chicago: University of Chicago Press, 21–60.
Green, Antony Dubach. 2001. The tense–lax distinction in English vowels and the role of parochial and analogical constraints. *Linguistics in Potsdam*, 15: 32–57.
—— 2005. Phonology limited. Manuscript. University of Potsdam.
Grimshaw, Jane. 1997. Projection, heads and optimality. *Linguistic Inquiry*, 28: 373–422.
Hahn, Reinhard. 1991. *Spoken Uyghur*. Seattle: University of Washington Press.
—— 1992. Modern Uyghur y ∼ r-insertion: Nativization through analogical extension. *Acta Linguistica Hafniensia*, 24: 77–96. Copenhagen: C. A. Reitzel.
Hale, Ken 1985. A note on Winnebago metrical structure, *International Journal of American Linguistics*, 51: 427–9.
—— and White Eagle, Josie. 1980. A preliminary metrical account of Winnebago accent. *International Journal of American Linguistics*, 46: 117–32.
Hale, M. 2000. Marshallese phonology, the phonetics-phonology interface and historical linguistics. *The Linguistic Review*, 17: 241–57.
—— 2007. *Historical Linguistics: Theory and Method*. Cambridge: Blackwell.
—— and Reiss, C. 1998. Formal and empirical arguments concerning phonological acquisition. *Linguistic Inquiry*, 29: 656–83.
—— —— 2000*a*. Substance abuse and dysfunctionalism: Current trends in phonology. *Linguistic Inquiry*, 31: 157–69.
—— —— 2000*b*. Phonology as cognition. In N. Burton-Roberts, Philip Carr, and Gerry Docherty (eds.), *Phonological Knowledge*. Oxford: Oxford University Press.
—— 2003. The subset principle in phonology: Why the tabula can't be rasa. *Journal of Linguistics*, 39: 219–44.
Halle, Morris. 1959. *The Sound Pattern of Russian*. The Hague: Mouton.
—— 1975. Confessio grammatici. *Language*, 51: 525–35.
—— 1987. Why phonological strata should not include affixation. Manuscript, MIT.
—— 1995. Feature geometry and feature spreading. *Linguistic Inquiry*, 26: 1–46.
—— and Idsardi, William. 1995. General properties of stress and metrical structure. In J. Goldsmith (ed.), *The Handbook of Phonological Theory*. Oxford: Blackwell, 403–43.
—— —— 1997. r, hypercorrection and the Elsewhere Condition. In Iggy Roca (ed.), *Derivations and Constraints in Phonology*. Oxford: Oxford University Press, 331–48.
—— and Marantz, Alec. 1993. Distributed morphology and the pieces of inflection. In Kenneth Hale and S. Jay Keyser (eds.), *The View from Building 20*. Cambridge: MIT Press, 111–76.
—— and Vergnaud, Jean-Roger. 1987. *An Essay on Stress*. Cambridge, MA: MIT Press.

Halle, Morris, Vaux, Bert, and Wolfe, Andrew. 2000. On feature spreading and the representation of place of articulation. *Linguistic Inquiry*, 31: 387–444.

—— O'Neil, Wayne, and Vergnaud, Jean-Roger. 1993. Metrical coherence in Old English without the Germanic foot. *Linguistic Inquiry*, 24.3: 529–39.

Hammond, Michael. 1994. An OT account of variability in Walmatjari stress. Manuscript. University of Arizona, Tucson. ROA-20.

—— 1999. *The Phonology of English: A Prosodic Optimality-Theoretic Approach*. Oxford: Oxford University Press.

Han, Eunjoo. 1994. Prosodic structure in compounds. Ph.D. dissertation. Stanford University.

Hanson, Kristin. 1992. Resolution in modern meters. Doctoral dissertation. Stanford, CA: Stanford University.

—— and Kiparsky Paul 1996. A theory of metrical choice. *Language*, 72: 287–335.

Harling-Kranck, Gunilla. 1998. *Från Pyttis till Nedervettil*. Helsingfors.

Harnad, Stevan (ed.). 1987. *Categorical Perception: The Groundwork of Cognition*. Cambridge: Cambridge University Press.

Harrison, Sheldon, with the assistance of Salich Albert. 1976. *Mokilese Reference Grammar*. Honolulu: University Press of Hawaii.

Hayes, Bruce. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.

—— 1996. Phonetically driven phonology: The role of Optimality Theory and inductive grounding. Milwaukee Conference on Formalism and Functionalism in Linguistics.

—— 2000. Gradient well-formedness in Optimality Theory. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford: Oxford University Press, 88–120.

—— 2004. Phonological acquisition in Optimality Theory: The early stages. In Rene Kager, Joe Pater, and Wim Zonneveld (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge: Cambridge University Press, 1–53.

—— and Albright, Adam. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90: 119–61.

—— and MacEachern, Margaret. 1998. Quatrain form in English folk verse. *Language*, 74: 473–507.

Horwood, Graham. 2000. A DIFFerent approach to Optional Optima. Manuscript. Rutgers University HUMDRUM I. <http://www.eden.rutgers.edu/~gvh/diff.pdf>.

Howard, Irwin. 1972. A directional theory of rule application in phonology. Doctoral dissertation, MIT.

Huldén, Lars. 1957. *Verbböjningen i Österbottens svenska folkmål*, 1. Helsingfors.

Hyman, Larry. 2000. The limits of phonetic determinism in phonology: *NC revisited. Talk presented at the LSA, Los Angeles.

—— and VanBik, Kenneth. 2002. Tone and syllable structure of the Hakha (Lai-Chin) noun. *Proceedings of the 28th Annual Berkeley Linguistic Society Meeting*, Special Session on Southeast Asian Linguistics, 15–28.

Idsardi, William. 1992. The Computation of Prosody. Doctoral dissertation, MIT.

—— 1997. Sympathy creates chaos. Manuscript, University of Delaware.

—— 1998. Tiberian Hebrew spirantization and phonological derivations. *Linguistic Inquiry*, 29: 37–73.

—— 2000. Clarifying opacity. *The Linguistic Review* 17.2–4: 337–50.

—— 2002. Further opacity issues: Spontaneors L2 Opacity. *Proceedings of the 2002 Linguistic Society of Korea International Summer Conference*, vol. II, 259–65.

—— 2006. A simple proof that Optimality Theory is computationally intractable. *Linguistic Inquiry*, 37: 271–5.

—— 2008. Calculating metrical structure. In Eric Raimy and Charles Cairns (eds.), *Contemporary Views on Architecture and Representations in Phonological Theory*. Cambridge, MA: MIT Press.

—— and Kim, Sun-Hoi. 2000. On Syllable-based multiple opacities. In Youngjun Jang and Jeong-Seok Kim (eds.), *International Proceedings of the 2000 Workshop on Generative Grammar*. Seoul: The Korean Generative Grammar Circle, 61–74.

Inkelas, Sharon. 1989. Prosodic Constituency in the Lexicon. Ph.D. dissertation, Stanford University. Published by Garland Press, 1990.

—— 1996. Archiphonemic underspecification. Manuscript. University of California, Berkeley.

—— and Orgun, Orhan. 1995. Level ordering and economy in the lexical phonology of Turkish. *Language*, 71: 763–93.

Isac, Daniela. 2002. Partitivity in the syntax and semantics of else. Manuscript. Concordia University, Montréal.

Itkonen, Terho. 1965. Über einige Sandhi-Erscheinungen im Finnmarklappischen, Finnischen und Finnlandschwedischen. *Finnisch-Ugrische Forschungen*, 35: 218–63.

Itō, Junko. 1986. Syllable theory in prosodic morphology. Doctoral dissertation, University of Massachusetts, Amherst.

—— and Mester, R. A. 1997. Sympathy Theory and German Truncations. *On'in kenkyuu Phonological Studies*, ed. Phonological Society of Japan, Katakusha, Tokyo, 1998, 51–66.

—— —— 1998. Sympathy theory and German truncations. Manuscript. Santa Cruz. (ROA 211.)

—— —— 1999. On the sources of opacity in OT: Coda processes in German. Manuscript. Santa Cruz. (ROA 347.)

—— —— 2003*a*. On the sources of opacity in OT: Coda processes in German. In Caroline Féry and Ruben van de Vijver (eds.), *The Syllable in Optimality Theory*. Cambridge: Cambridge University Press, 271–303.

—— —— 2003*b*. Lexical and postlexical phonology in Optimality Theory: Evidence from Japanese. In G. Fanselow and C. Féry (eds.), *Linguistische Berichte. Sonderheft 11: Resolving Conflicts in Grammars*, 183–207.

Jackendoff, Ray. 1990. *Semantic Structures*. Cambridge, MA: MIT Press.

Jensen, J. T. 1977. *Yapese Reference Grammar*. Honolulu: University Press of Hawaii.

Johnson, C. Douglas. 1972. *Formal Aspects of Phonological Description*. The Hague: Mouton.

Johnson, Mark. 1984. A discovery procedure for certain phonological rules. 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics.

Kager, R. 1997. Rhythmic vowel deletion in OT. In Iggy Roca (ed.), *Derivations and constraints in Phonology*. Oxford: Clarendon.

—— 1999. *Optimality Theory*. Cambridge: Cambridge University Press.

Kaplan, Ronald. 1995. Three seductions of computational psycholinguistics. In Mary Dalrymple, Ronald Kaplan, John Maxwell III, and Annie Zaenen (eds.), *Formal Issues in Lexical-Functional Grammar*. Palo Alto: CSLI. (Originally published in 1987.)

Karttunen, Lauri. 1993. Finite state constraints. In John Goldsmith (ed.), *The Last Phonological Rule*. Chicago: University of Chicago Press, 173–94.

—— 1998. The proper treatment of Optimality in Computational Phonology. In the *Proceedings of the International Workshop on Finite-State Methods in Natural Language Processing*, June 29-July 1, 1998:. Bilkent University, Ankara, Turkey, 1–12.

Kaufmann, F. 1887. Zum germanischen Consonantismus. *Beitrage zur Geschicte der Deutschen Sprache and Literatur*, 12: 511–47.

Kautz, Henry, and Selman, Bart. 1991. Hard problems for simple default logic. *Artificial Intelligence Journal*, 49: 243–79.

Kawasaki, Haruko. 1982. An acoustical basis for universal constraints on sound sequences. Doctoral dissertation, University of California, Berkeley.

—— 1992. An acoustical basis for universal phonotactic constraints. *Language and Speech*, 35.1–2: 73–86.

Kaye, Jonathan. 1989. *Phonology: A Cognitive View*. Hillsdale, NJ: Lawrence Erlbaum.

—— Lowenstamm, Jean, and Vergnaud, Jean-Roger. 1985. The internal structure of phonological elements: A theory of charm and government. *Phonology Yearbook*, 2: 303–28.

—— —— —— 1990. Constituent structure and government in phonology. *Phonology*, 7: 193–231.

Kean, Mary-Louise. 1975. The theory of markedness in Generative Grammar. Doctoral dissertation. MIT.

Keating, P. A. 1988. Underspecification in phonetics. *Phonology*, 5.2: 275–92.

Keer, Edward W. 1999. Geminates, the Ocp and the Nature of CON. Ph.D. dissertation, Rutgers University. New Brunswick, New Jersey.

Keller, Frank. 1998. Gradient grammaticality as an effect of selective constraint reranking. In M. C. Gruber, D. Higgins, K. Olson, and T. Wysocki (eds.), *Papers from the 34th Meeting of the Chicago Linguistic Society*, vol. 2. Chicago: University of Chicago. 95–109.

Kempson, R. 1988. On the grammar-cognition interface. In R. Kempson (ed.), *Mental representations*: *The Interface between Language and Reality*. Cambridge and New York: Cambridge University Press, 1988.

Kenstowicz, Michael. 1994*a*. Cyclic vs. noncyclic constraint evaluation. *Phonology*, 12: 397–436.
—— 1994*b*. *Phonology in Generative Grammar*. Oxford: Blackwell.
—— and Kisseberth, Charles. 1970. Rule ordering and the asymmetry hypothesis. *Papers from the 6th Regional Meeting*, Chicago Linguistic Society, 504–19.
—— —— 1979. *Generative Phonology: Description and Theory*. New York: Academic Press.
—— and Sohn, Hyang-Sook. 2000. Accentual adaptation in North Kyungsang Korean. Manuscript. MIT. To appear in M. Kenstowicz (ed.), *Ken Hale: A Life in Language*. Cambridge, MA: MIT Press.
Keyser, Jay and O'Neil, Wayne. 1985. *Rule Generalization and Optionality in Language Change*. Dordrecht: Foris.
—— and Stevens, Kenneth. 2006. Enhancement and overlap in the speech chain. *Language*, 82.1: 33–63.
Kiparsky, Paul. 1972. Explanation in Phonology. In S. Peters (ed.), *Goals of Linguistic Theory*. Englewood, NJ: Prentice-Hall, 189–227.
—— 1973. Phonological representations. In Osamu Fujimura (ed.), *Three Dimensions of Linguistic Theory*. Tokyo: Tokyo Institute for Advanced Studies of Language, 3–136.
—— 1982*a*. *Explanation in Phonology*. Dordrecht: Foris.
—— 1982*b*. From cyclic phonology to lexical phonology. In Harry van der Hulst and Norval Smith (eds.), *The Structure of Phonological Representations*. Dordrecht: Foris.
—— 1982*c*. Lexical phonology and morphology. In I. S. Yang (ed.), *Linguistics in the Morning Calm*. Seoul: Hanshin, 3–91.
—— 1984. On the lexical phonology of Icelandic. In C.-C. Elert, I. Johnson, and Eva Stangert (eds.), *Nordic Prosody*, III. Stockholm: Almquist & Wiksell, 135–64.
—— 1995. The phonological basis of sound change. In John Goldsmith (ed.), *The Handbook of Phonological Theory*. Cambridge, MA: Blackwell, 640–70.
—— 1997. LP and OT. Handout from LSA Summer Linguistic Institute, Cornell University.
—— 1999. Paradigm effects and opacity. Manuscript. Stanford University.
—— 2000. Opacity and cyclicity. *The Linguistic Review*, 17.2–4: 351–66.
—— 2003. Syllables and moras in Arabic. In C. Fery and R. Vijver (eds.), *The Syllable in Optimality Theory*. Cambridge: Cambridge University Press.
—— 2005. Where Stochastic OT fails: A discrete model of metrical variation. *Proceedings of BLS*.
—— Forthcoming. *Paradigm Effects and Opacity*. Stanford: CSLI.
Kirchner, Robert. 1966. Synchronic chain shifts in Optimality Theory. *Linguistic Inquiry*, 27: 341–50.
—— 1996. Synchronic chain shifts in Optimality Theory. *Linguistic Inquiry* 27.2: 341–50.
—— 2001. Review of April McMahon (2000), *Change, Chance, and Optimality*. *Phonology*, 18.3: 427–33.

Kisseberth, Charles. 1970. On the functional unity of phonological rules. *Linguistic Inquiry*, 1: 291–306.

Kissock, Madelyn, Hale, Mark, and Reiss, Charles. 1998. What is output? Output-output correspondence in OT phonology, (with M. Hale and C. Reiss) in *Proceedings of the West Coast Conference on Formal Linguistics*, XVI. CSLI Publications/Cambridge University Press, 223–36.

——— ——— ——— 2000. Evaluating the empirical basis for output-output correspondence. In M. Juge and J. Moxley (eds.), *Proceedings of the Berkeley Linguistic Society*, Berkeley Linguistics Society, 137–47.

Kitto, Catherine and Lacy, Paul de. 1999. Correspondence and epenthetic quality. In Catherine Kitto and Carolyn Smallwood (eds.), *Proceedings of AFLA VI*. Toronto: Toronto Working Papers in Linguistics, 181–200. Also ROA 337.

Kolinsky, Regine. 1994. Intermediate representations in spoken word recognition: Evidence from word illusions. *Journal of Memory and Language*, 34: 19–40.

Kornai, Andras. 2006. Is OT NP-hard? ROA 838-0606.

Koskeniemi, Kimmo. 1983. Two-level morphology. Publication 11, Department of Linguistics, University of Helsinki.

Kraehenmann, Astrid. 2001*a*. Swiss German stops: Geminates all over the word. *Phonology*, 18: 109–46.

——— 2001*b*. Quantity and prosodic asymmetries in Alemannic: synchronic and diachronic perspectives. Ph.D. dissertation, University of Konstanz.

Lacharité, Darlene and Paradis, Carole. 1993. The emergence of constraints in generative phonology and a comparison of three current constraint-based models. *The Canadian Journal of Linguistics*, 38.2: 127–303.

Lacy, Paul de. 1998. Sympathetic stress. Manuscript. University of Massachusetts-Amherst. (ROA 294).

Lakoff, G. 1993. Cognitive phonology. In J. Goldsmith (ed.), *The Last Phonological Rule*. Chicago: University of Chicago Press, 117–45.

Legendre, Géraldine, Sorace, Antonella, and Smolensky, Paul. (2006). The Optimality Theory–Harmonic Grammar connection. In Smolensky and Legendre (2006), 903–66.

Lerdahl, Fred and Jackendoff, Ray. 1983. *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Levi, Susannah. 2000. Modern Hebrew: A challenge for Sympathy. *University of Washington Papers in Linguistics*, 19: 1–14.

Levin, Julliete. 1985. A metrical theory of syllabicity. Doctoral dissertation, MIT.

Liberman, Mark. 1975. The intonational system of English. Doctoral dissertation, MIT.

Liphola, M. 1999. Stress-dependent vowel reduction in Shimakonde. Manuscript. Ohio State University.

——— 2001. Aspects of the phonology and morphology of Shimakonde. Doctoral dissertation, Ohio State University.

List, Christian and Harbour, Daniel. 2001. Optimality Theory and the problem of constraint aggregation. In *MIT Working Papers in Linguistics and Philosophy*, 1: *The Linguistics/Philosophy Interface*.

Local, J. and Coleman, J. 1994. The no crossing constraint. *Linguistics and Philosophy*, 14: 295–338.

Lombardi, Linda. 1997. Coronal epenthesis and markedness. *University of Maryland Working Papers in Linguistics* 5: 156–75. Also available at ROA-245.

—— 2001. Why place and voice are different: constraint-specific alternations in Optimality Theory. In Linda Lombardi (ed.), *Segmental Phonology in Optimality Theory*. Cambridge: Cambridge University Press, 13–45.

Lubowicz, Anna. 1998. Derived environment effects in OT. Manuscript. University of Massachussetts (ROA 239).

Lynch, J. 1978. A grammar of Lenakel. *Pacific Linguistics*, Series B, 55. Australian National University, Canberra.

Lyons, John. 1970. *Noam Chomsky*. New York: Viking Press.

Marantz, Alec. 1982. Re reduplication. *Linguistic Inquiry*, 13.3: 435–82.

Marlett, S. and Stemberger, J. 1983. Empty consonants in Seri. *Linguistic Inquiry*, 14: 617–39.

McCarthy, John. 1979. On Stress and Syllabification. *Linguistic Inquiry*, 10: 443–65.

—— 1986. Ocp effects: Gemination and antigemination. *Linguistic Inquiry*, 17: 207–63.

—— 1988. Feature geometry and dependency: A review. *Phonetica*, 45: 84–108.

—— 1996. Remarks on phonological opacity in Optimality Theory. In Jacqueline Lecarme, Jean Lowenstamm, and Ur Shlonsky (eds.), *Studies in Afroasiatic Grammar. Papers from the Second Conference on Afroasiatic Linguistics*. The Hague: Holland Academic Graphics, 215–43.

—— 1997*a*. Process-specific constraints in Optimality Theory. *Linguistic Inquiry*, 28.2: 231–51.

—— 1997*b*. Sympathy and phonological opacity. Handout of lecture given at MIT, April 4.

—— 1998. Sympathy and phonological opacity. ROA 252-0398; published in 2000 in *Phonology*, 16: 331–401.

—— 1999*a*. Introductory OT on CD-ROM. Amherst, MA: Graduate Linguistic Students' Association.

—— 1999*b*. Sympathy and phonological opacity. *Phonology*, 16: 331–99.

—— 1999*c*. Serialism, OT, and the Duke-of-York Gambit. Rutgers Optimality Archive.

—— 1999*d*. Sympathy, cumulativity, and the Duke-of-York gambit. Manuscript. University of Massachusetts.

—— 2000. The Prosody of Phase in Rotuman. Natural Language and Linguistic Theory 18, 147–97.

—— 2002*a*. A thematic guide to Optimality Theory. Cambridge: Cambridge University Press.

McCarthy, John. 2002*b*. On targeted constraints and cluster simplification. *Phonology*, 19.2: 273–92.

—— 2003*a*. Sympathy, cumulativity, and the Duke-of-York gambit. In Caroline Féry and Ruben van de Vijver (eds.), *The Syllable in Optimality Theory*. Cambridge: Cambridge University Press, 23–76.

—— 2003*b*. OT constraints are categorical. *Phonology*, 20: 75–138.

—— 2003*c*. Comparative markedness. *Theoretical Linguistics*, 29: 1–51.

—— 2005*a*. Taking a free ride in morphophonemic learning. *Catalan Journal of Linguistics*, 4.

—— 2005*b*. Optimal paradigms. In Laura Downing, Tracy Alan Hall, and Renate Raffelsiefen (eds.), *Paradigms in Phonological Theory*. Oxford: Oxford University Press, 170–210.

—— 2006. Candidates and derivations in Optimality Theory. ROA 823-0506.

—— 2007. *Hidden Generalizations: Phonological Opacity in Optimality Theory*. London: Equinox.

—— and Prince, Alan. 1993. Prosodic morphology I: constraint interaction and satisfaction. Manuscript. University of Massachussets, Amherst and Rutgers University.

—— —— 1994*a*. The emergence of the unmarked: Optimality in prosodic morphology. In *NELS*, 24. Amherst: GLSA, University of Massachusetts, 333–79.

—— —— 1994*b*. Two lectures on prosodic morphology. ROA 59-0000.

—— —— 1995*a*. Faithfulness and reduplicative identity. In *University of Massachusetts Occasional Papers in Linguistics, 18: Papers in Optimality Theory*. Amherst: GLSA, University of Massachusetts, 77–136.

—— —— 1995*b*. Prosodic morphology. In John Goldsmith (ed.), *The Handbook of Phonological Theory*. Oxford: Blackwell.

—— —— 1999. Faithfulness and identity in prosodic morphology. In Rene Kager, Harry Van Der Hulst, and Wim Zonneveld (eds.), *The Prosody–Morphology Interface*. Cambridge: Cambridge University Press, 218–309.

—— and Wolf, Matthew. 2005. Less than zero: Correspondence and the null output. ROA 722-0305.

McMahon, April. 1998. Optimality and optimism. Paper presented to the Linguistic Circle, University of Edinburgh. <http://www.ling.ed.ac.uk/discussion/lcircle/abs/mcmahon.abstract>.

Mester, Armin. 1992. Some remarks on Tongan stress. Manuscript. University of Santa Cruz.

Miner, Kenneth. 1979. Dorsey's Law in Winnebago-Chiwere and Winnebago accent, *International Journal of American Linguistics*, 45: 25–33.

—— 1981. Metrics, or Winnebago made harder. *International Journal of American Linguistics*, 47: 340–2.

—— 1989. Winnebago accent: The rest of the data. *Anthropological Linguistics*, 31: 148–72.

Mitchell, T. F. 1960. Prominence and syllabification in Arabic. *Bulletin of the School of Oriental and African Studies*, 23: 369–89. Reprinted in T. F. Mitchell. 1975. *Principles of Firthian Linguistics*. London: Longmans.

Mithun, Marianne and Basri, Hasan. 1986. The phonology of Selayarese. *Oceanic Linguistics*, 25: 210–54.

Mohanan, Karuvannur Puthanveettil. 2000. The theoretical substance of the optimality formalism. *The Linguistic Review*, 17: 143–66.

Moravcsik, Edith. 1978. Reduplicative constructions. In Joseph Greenberg (ed.), *Universals of Human Language*, vol. 3. Stanford: Stanford University Press, 297–334.

Moreton, Elliott and Smolensky, Paul. 2002. Typological consequences of local constraint conjunction. In L. Mikkelsen and C. Potts (eds.), *Proceedings of the West Coast Conference on Formal Linguistics*, 21. Cambridge, MA: Cascadilla Press, 306–19.

Myers, Scott. 1991. Persistent rules. *Linguistic Inquiry*, 22: 315–44.

—— 1997. OCP effects in optimality theory. *NLLT*, 15.4: 847–92.

—— 2002. Gaps in factorial typology: The case of voicing in consonant clusters. Manuscript. University of Texas. <uts.cc.utexas.edu/∼smyers/voicing.pdf>.

Nevins, Andrew. 2007. Review of Andrea Calabrese, *Markedness and Economy in a Derivational Model of Phonology. Journal of Linguistics*, 43: 223–67.

—— and Vaux, Bert. 2003. Metalinguistic, shmetalinguistic: The phonology of shm-reduplication. *CLS*, 39: 702–21.

—— —— 2004. Consonant harmony in Karaim. *MIT Working Papers in Linguistics*, 46: 175–94.

—— and Endress, Ansgar. 2007. The edge of order: Analytic bias in ludlings. Jeremy Rau, Keith Plaster, Patrick Liu, and Yaroslav Gorbachov (eds.), *Harvard Working Papers in Linguistics*, 12, 43–53.

Newton, Brian. 1972. *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology*. Cambridge: Cambridge University Press.

Ní Chiosáin, M. and Padgett, Jaye. 1993. Inherent V-Place. *UC Santa Cruz Working Papers*. Linguistics Research Center, University of California, Santa Cruz.

Noyer, Rolf. 1994. Palatalization and vowel place in San Mateo Huave: The competition of syntagmatic and paradigmatic well-formedness. Paper presented at the LSA Annual meeting.

Núñez-Cedeño, Rafael. 1988. Structure-preserving properties of an epenthetic rule in Spanish. In David Birdsong and Jean-Pierre Montreuil (eds.), *Advances in Romance Linguistics*. Dordrecht: Foris, 319–36.

Odden, D. 1986. On the role of the obligatory contour principle in phonological theory. *Language*, 62: 353–83.

—— 1987. Kimatuumbi phrasal phonology. *Phonology Yearbook*, 4: 13–36.

—— 1988. Antiantigemination and the Ocp. *Linguistic Inquiry*, 19: 451–75.

—— 1990. Syntax, lexical rules and postlexical rules in Kimatuumbi. In S. Inkelas and D. Zee (eds.), *Proceedings of the Phonology ∼ Syntax Connection Conference*. Chicago: University of Chicago Press, 259–77.

—— 1995. Onsetless syllables in Kikerewe. *OSUWPL*, 47: 89–110.

—— 1996. *The Phonology and Morphology of Kimatuumbi*. Oxford: Clarendon.

Odden, D. 1998. Verbal tone melodies in Kikerewe. In T. Hinnebusch and I. Maddieson (eds.), *Theoretical Approaches to African Linguistics, II*. New Brunswick: Red Sea Press, 177–84.

—— 2000. Opacity and Ordering: H-deletion in Kikerewe. *Linguistic Review*, 17: 323–35.

—— 2005. *Introducing Phonology*. Cambridge: Cambridge University Press.

Ohala, John. 1971. The role of physiological and acoustic models in explaining the direction of sound change. Project on Linguistic Analysis Reports (Berkeley) 15: 25–40.

—— 1972. How to represent natural sound patterns. Project on Linguistic Analysis (Berkeley) 16: 40–57.

—— 1975. Phonetic explanations for nasal sound patterns. In C. A. Ferguson, L. M. Hyman, and J. J. Ohala (eds.), *Nasálfest: Papers from a Symposium on Nasals and Nasalization*. Stanford: Language Universals Project, 289–316.

—— 1981. The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, and M. F. Miller (eds.), *Papers from the Parasession on Language and Behavior*. Chicago: Chicago Linguistic Society, 178–203.

—— 1990. The phonetics and phonology of aspects of assimilation. In J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology, I: Between the Grammar and Physics of Speech*. Cambridge: Cambridge University Press.

—— 2005. Phonetic explanations for sound patterns. Implications for grammars of competence. In W. J. Hardcastle and J. M. Beck (eds.), *A Figure of Speech. A Festschrift for John Laver*. London: Erlbaum, 23–38.

—— and Lorentz, James. 1977. The story of [w]: an exercise in the phonetic explanation for sound patterns. *Proceedings of the Annual Meeting of the Berkeley Linguistic Society*, 3: 577–99.

Oostendorp, Marc van. 2003. Comparative markedness and containment. *Theoretical Linguistics*, 29.1/2: 65–76.

—— 2005. Derived environment effects and consistency of Exponence. Workshop on Freedom of Analysis, University of Tromsø, September 1–2, 2005.

Orgun, C. Orhan. 1993. Monotonic cyclicity and Optimality Theory. Talk delivered at Stanford University. <www-csli.stanford.edu/Archive/calendar/1993–94/msg00008.html>.

—— 1996*a*. Monotonic cyclicity. Manuscript. University of California-Berkeley. ROA 123.

—— 1996*b*. OT versus two-level phonology: Limits of faithfulness. Stanford Phonology Workshop. Abstract <www-csli.stanford.edu/Archive/calendar/1995–96/msg00027.html+koskenniemi+ot&hl=en&ct=clnk&cd=2&gl=uk> accessed on April 26, 2007.

—— 1996*c*. Correspondence and identity constraints in two-level Optimality Theory. In J. Camacho, L. Choueiri, and M. Watanabe (eds.), *WCCFL*, 14. CSLI: Stanford, 399–413.

Orgun, C. Orhan. 1996*d*. Sign-based phonology and morphology, with special attention to optimality. Ph.D. dissertation. University of California, Berkeley.

—— and Sprouse, Ronald L. 1999. From MPARSE to CONTROL: Deriving ungrammaticality. Phonology, 16: 191–224.

Ouden, Dirk-Bart den. 2001. Review of "The Derivational Residue in Phonological Optimality Theory [Linguistics Today]" by Ben Hermans and Mare Van Oostendorp (eds.). *Studies in Language*, 25.3: 651–7.

—— 2002. Review of Hermans and van Oostendorp 1999. Fundamentals of Language: 651–7.

Padgett, Jaye. 2003. Contrast and post-velar fronting in Russian. *Natural Language and Linguistic Theory*, 21: 39–87.

Paradis, Carole. 1988*a*. On constraints and repair strategies. *The Linguistic Review*, 6: 71–97.

Paradis, Carole. 1988*b*. Towards a theory of constraint violations. *McGill Working Papers in Linguistics*, 5: 1–43.

Parkinson, F. 1996. The representation of vowel height in phonology. Doctoral dissertation, Ohio State University.

Pater, Joe. 1999. Austronesian nasal substitution and other NC effects. In The prosody-morphology interface, Harry van der Hulst, René Kager, and Wim Zonneveld (eds.), The Prosody Morphology Interface Cambridge: Cambridge University Press, 310–43.

—— 2000. Nonuniformity in English stress: The role of ranked and lexically specific constraints. Phonology, 17.2: 237–74.

—— 2001. Austronesian nasal substitution revisited. In Linda Lombardi (ed.), *Segmental Phonology in Optimality Theory: Constraints and Representations*. Cambridge: Cambridge University Press, 159–182.

—— 2003. Balantak metathesis and theories of possible repair in Optimality Theory Manuscript, University of Massachusetts, Amherst.

—— 2005. Non-convergence in the GLA and variation in the CDA. ROA 780–1005.

—— 2007*a*. The power of weighted constraints. Colloquium presentation, Johns Hopkins University, February 8, 2007. Abstract accessed at <http://people.umass.edu/pater/pater-hopkins.pdf> on April 25, 2007.

—— 2007*b*. Local harmonic serialism. Workshop presentation at CASTL, Tromsø, March 27–8.

—— and John McCarthy. 2004. Targeted constraints and opacity. Handout from Ling 730, retrieved from courses.umass.edu/ling730/tcotandopacity.pdf on 26 May 2007.

Peng, Long. 2002. Local constraint conjunction and Kikuyu consonant mutation. Penn Linguistics Colloquium 26.

Perlmutter, David. 1988. The split morphology hypothesis: Evidence from Yiddish. In Michael Hammond and Michael Noonan, (eds.), *Theoretical Morphology*. San Diego: Academic Press.

Pesetsky, David. 1997. Optimality Theory and syntax: Movement and pronunciation. In Diana Archangeli: and D. Terence Langendoen (eds.), *Optimality Theory: An Overview*. Oxford: Blackwell, 134–70.

——1998. Some optimality principles of sentence pronunciation. In Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky (eds.), *Is the Best Good Enough? Optimality and Competition in Syntax*, Cambridge, MA: MIT Press, 337–83.

Phelps, E. 1973. Tonkawa, Sundanese, and Kasem: Some problems in generative phonology. Doctoral dissertation, University of Washington.

——and M. Brame. 1974. On local ordering of rules in Sanskrit. *Linguistic Inquiry*, 4: 387–400.

Piattelli-Palmarini, Massimo. 1994. *Inevitable Illusions: How Mistakes of Reason Rule our Minds*. New York: Wiley.

Picanco, Gessiane. 2006. Mundurukú: Phonetics, phonology, synchrony, diachrony. Doctoral dissertation, University of British Columbia.

Piggott, Glyne L. 1995. Epenthesis and syllable weight. *Natural Language and Linguistic Theory*, 13: 283–326.

Pinker, S. 1984/96. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

——and Prince, Alan. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28: 73–193.

Pipping, Hugo. 1892–7. Om det bildade uttalet av svenska sproket [sic] i Finland. Nystavaren. (Non vidi.)

Poletto, R. 1998. *Topics in Runyankore phonology*. Doctoral dissertation, Ohio State University.

Poliquin, Gabriel. 2006. Canadian French vowel harmony. Doctoral dissertation, Harvard University.

Postal, Paul. 2004. *Skeptical Linguistic Essays*. Oxford: Oxford University Press.

Potts, Christopher and Pullum, Geoffrey. 2002. Model theory and the content of OT constraints. *Phonology*, 19: 361–93.

Prince, Alan. 1976. "Applying" stress. Manuscript. University of Massachusetts.

——and Smolensky, Paul. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical Report RUCCS, Rutgers University, New Brunswick, NJ. Published in 2004 by Blackwell.

————2002. Optimality Theory: Constraint interaction in generative grammar. <http://roa.rutgers.edu/files/537–0802/537–0802-PRINCE-0–0.PDF>.

————2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell.

Pullum, Geoff. 1976. The Duke of York gambit. *Journal of Linguistics*, 12: 83–102.

Pycha, Anne, Nowak, Pawel, Shosted, Ryan, and Shin, Eurie. 2003. Phonological rule-learning and its implications for a theory of vowel harmony. In M. Tsujimura and G. Garding (eds.), WCCFL Proceedings, 22 Somerville, MA: Cascadilla Press, 101–14.

Pylyshyn, Zenon W. 1973. On the role of competence theories in cognitive psychology. *Journal of Psycholinguistic Research*, 2: 21–50.

——1984. Computation and cognition: Toward a foundation for cognitive science. Cambridge, MA: MIT Press.

Pylyshyn, Zenon W. 2003. Seeing and Visualizing: It's not what you think. Cambridge, MA: MIT Press.

Quine, Willard V. 1972. Methodological reflections on current linguistic theory. In G. Harman and D. Davidson (eds.), *Semantics of Natural Language*. New York: Humanities Press.

Raimy, Eric. 1999. Representing reduplication. Doctoral dissertation, University of Delaware.

—— 2000*a*. *The phonology and morphology of reduplication*. Berlin: Mouton de Gruyter.

—— 2000*b*. Remarks on backcopying. *Linguistic Inquiry*, 31: 541–52.

—— and Idsardi, William. 1997. A minimalist approach to reduplication in OT. In Kiyomi Kusumoto (ed.), *NELS 27*. Amherst: GLSA, University of Massachusetts. 369–83.

Reiss, Charles. 1995. A theory of assimilation, with special reference to Old Icelandic phonology. Ph.D. dissertation, Department of Linguistics, Harvard University.

—— 1999. Acquisition and Post-OT phonology. To appear in B. Vaux and M. Halle (eds.), *Phonology*. 2000.

—— 2000. Optimality Theory from a cognitive science perspective. *The Linguistic Review*, 17: 291–301.

—— 2003*a*. Quantification in structural descriptions: Attested and unattested patterns. *The Linguistic Review*, 20: 305–38.

—— 2003*b*. Deriving the feature-filling/feature-changing contrast: An application to Hungarian vowel harmony. *Linguistic Inquiry*, 34: 199–224.

Rennison, J. 2000. OT and TO. *The Linguistic Review*, 17: 135–41.

Reuter, Mikael. 1982. Kvantitet i helsingforssvenskan. Manuscript. Helsingfors.

—— 1986. Kortstavighet i helsingforssvenskan. *Festskrift till Carl-Erik Thors*. Helsingfors: Svenska litteratursällskapet.

Riad, Tomas. 1992. *Structures in Germanic Prosody*. Stockholm: Stockholm University.

Rice, Curt. 2005. Optimal gaps in optimal paradigms. *Catalan Journal of Linguistics*, 4: 155–70.

—— 2007. Gaps and repairs at the phonology–morphology interface. *Journal of Linguistics* 43: 197–221.

Riggle, Jason and Wilson, Colin. 2006. Local Optionality. In Leah Bateman and Cherlon Ussery (eds.), *Proceedings of the Thirty-Fifth Annual Meeting of the North East Linguistic Society*, vol. 2, Charleston, SC: BookSurge Publishing.

Robins, R. H. and Waterson, N. 1952. Notes on the phonetics of the Georgian word. *Bulletin of the School of Oriental and African Studies*, 15: 55–72.

Rose, Sharon. 2000. Rethinking geminates, long-distance geminates, and the Ocp. *Linguistic Inquiry*, 31: 85–122.

Rubach, Jerzy. 1997. Extrasyllabic consonants in Polish: Derivational Optimality Theory. In I. Roca (ed.), *Derivations and Constraints in Phonology*. Oxford: Oxford University Press, 551–81.

—— 2000. Glide and glottal stop insertion in Slavic Languages: A DOT analysis. *Linguistic Inquiry*, 31: 271–317.

—— 2003. Duke-of-York derivations in Polish. *Linguistic Inquiry*, 34.4: 601–29.

—— 2004. Derivation in Optimality Theory: A reply to Burzio. *Linguistic Inquiry*, 35.4: 656–70.

Salting, Donald. 2005. The geometry of harmony: Evidence against targeted constraints. In Marc van Oostendorp and Jeroen van de Weijer (eds.), *The Internal Organization of Phonological Segments*, Berlin: Mouton de Gruyter.

Sampson, Geoffrey. 1975. One fact needs one explanation. *Lingua*, 36: 231–9.

Sapir, E. 1915. Notes on Judeo-German phonology. *The Jewish Quarterly Review*, n.s., 6. 231–66. Reprinted in David G. Mandelbaum (ed.), *Selected Writings of Edward Sapir in Language, Culture, and Personality*. Berkeley: University of California Press.

—— 1930. Southern Paiute, A Shoshonean language. *Proceedings of the American Academy of Arts and Sciences*, 65: 1–296.

Schein, Barry and Steriade, Donca. 1986. On geminates. *Linguistic Inquiry*, 17.4: 691–744.

Schütz, Albert. 1985. *The Fijian Language*. Honolulu: University of Hawaii Press.

Sebregts, Koen. 2001. English [r]-liaison: rule-based theories, Government Phonology and Optimality Theory. Master's thesis, Department of English, University of Leiden.

Seidenberg, Mark and MacDonald, Maryellen. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23: 569–88.

Selenius, Ebba. 1972. Västnyländsk ordaccent. Helsingfors: Svenska Litteratursällskapet i Finland.

Selkirk, Elisabeth. 1999. Morphologically governed Output-Output constraints in a noncyclic Optimality Theoretic grammar: Evidence from the Makassar languages. Talk give at Sophia University, Tokyo.

Sherwood, David. 1986. *Maliseet-Passamaquoddy Verb Morphology*. Ottawa, Ontario: National Museums of Canada.

Sherzer, Joel. 1970. Talking backwards in Cuna: The sociological reality of phonological descriptions. *Southwestern Journal of Anthropology*, 26: 343–53.

Shwayder, Kobey. 2007. A study of productivity and variation in Icelandic *u*-umlaut. Bachelor's thesis, Harvard University.

Singh, Rajendra. 1987. Well-formedness conditions and phonological theory. In Wolfgary Dresser (ed.), *Phonologica 1984*. Cambridge: Cambridge University Press, 273–86.

Sirk, Ülo. 1988. Towards the historical grammar of the South Sulawesi languages: possessive enclitics in the postvocalic position. In *Pacific Linguistics, A-79, Papers in Western Austronesian Linguistics*, 4: 283–302.

Smolensky, Paul. 1996. On the comprehension/production dilemma in child language. *Linguistic Inquiry*, 27: 720–31.

—— 1999. Grammar-based connectionist approaches to language. *Cognitive Science*, 23: 589–613.

Smolensky, Paul. and Legendre, Géraldine. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, *volume I*: *Cognitive Architecture*. Cambridge, MA: MIT Press.

Sohn, Ho-min and Bender, Byron W. 1973. *A Ulithian Grammar*. Pacific Linguistics, Series C—No. 27. Linguistic Circle of Canberra.

Sommerstein, Alan. 1974. On phonotactically motivated rules. *Journal of Linguistics*, 10: 71–94.

Spaelti, Philip. 1997. Dimensions of variation in multi-pattern reduplication. Ph.D. dissertation, University of California, Santa Cruz.

Sproat, Richard. 1985. On deriving the lexicon. Doctoral dissertation, MIT.

—— 1988. Bracketing paradoxes, cliticization, and other topics. In M. Everaert *et al.* (eds.), *Morphology and Modularity*. Dordrecht: Foris.

Sprouse, Ronald. 1997. A case for enriched inputs. Manuscript. University of California, Berkeley.

Stemberger, Joseph. 1996. The scope of the theory: Where does beyond lie? In Lisa McNair, Kora Singer, Lise M. Dobrin, and Michelle M. Aucon (eds.), *CLS 32: Papers from the Parasession on Theory and Data in Linguistics*. Chicago Linguistics Society. 139–64.

Steriade, Donca. 1988. Reduplication and syllable transfer in Sanskrit and elsewhere. *Phonology*, 5: 73–155.

—— 1990. Gestures and autosegments: Comments on Browman and Goldstein's paper. In J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology*. Cambridge: Cambridge University Press, 382–97.

—— 2000. Paradigm Uniformity and the phonetics/phonology boundary. In Janet Pierrehumbert and Michael Broe (eds.), *Papers in Laboratory Phonology*, Cambridge: Cambridge University Press, 313–35.

—— 2001. The phonology of perceptibility effects: The P-map and its consequences for constraint organization. Manuscript. MIT.

Stevens, Ken, Keyser, S. Jay, and Kawasaki, H. 1986. Towards a phonetic and phonological theory of redundant features. In J. S. Perkell and Dennis Klatt (eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum.

Streiff, Catharine. 1915. *Die Laute der Glarner Mundarten*. Frauenfeld: Huber.

Struijke, Caro. 2000. Why constraint conflict can disappear in reduplication. Manuscript. University of Maryland, College Park. ROA 373.

—— 2002. *Existential Faithfulness: A Study of Reduplicative TETU, Feature Movement, and Dissimilation*. Routledge Outstanding Dissertations in Linguistics. New York: Routledge.

Sundberg, Eva. 1993. *Dialekten i Ålands nordöstra skärgård*. Ekenäs: Ålands högskola.

Szabolcsi, Anna. 1988. Filters versus Combinators. In Istv'an Bodn'ar, Andr's M'at'e and L'aszlo P'olos (eds.), *Intensional Logic, History of Philosophy, and Methodology*. Budapest.

Taylor, S. 1969. *Koya: An Outline Grammar, Gommu Dialect*. University of California Publications in Linguistics, 54. Berkeley and Los Angeles: University of California Press.

Tenenbaum, Joshua. 1999. A Bayesian framework for concept learning. Doctoral dissertation, MIT.

—— and Griffiths, Thomas. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24: 629–41.

Tesar, Bruce and Smolensky, Paul. 1996. Learnability in Optimality Theory (long version). Technical report 96-3. Department of Cognitive Sciences, Johns Hopkins University, Baltimore. Available as Rutgers Optimality Archive 156.

—— —— 1998. Learnability in Optimality Theory. *Linguistic Inquiry*, 29: 229–68.

—— —— 2000. *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.

Toivonen, Ida. 2001. The phrase structure of non-projecting words. Ph.D. dissertation, Stanford University.

Uffmann, Christian. 2004. Re-evaluating transfer in SLA. Montreal dialogues, UQAM. Accessed on 8 June 2007 from <http://staff-www.uni-marburg.de/~uffmann/Montreal04-handout.pdf>.

Uhrbach, Amy. 1987. A formal analysis of reduplication and its interaction with phonological and morphological processes. Ph.D. dissertation, Department of Linguistics, University of Texas, Austin.

Urbanczyk, Suzanne. 1996. Patterns of reduplication in Lushootseed. Ph.D. dissertation, University of Massachusetts, Amherst.

—— 2001. *Patterns of Reduplication in Lushootseed*. Outstanding Dissertations in Linguistics. New York: Garland.

—— 2006. Reduplicative form and the root-affix asymmetry. *Natural Language and Linguistic Theory*, 24: 179–240.

Vaux, Bert. 1998. *The Phonology of Armenian*. Oxford: Clarendon Press.

—— 2001. Consonant insertion and hypercorrection. Paper presented at the LSA Annual Meeting, Washington, DC. [Manuscript available on request.]

—— 2003. Why the phonological component must be serial and rule-based. Paper presented at the Annual Meeting of the Linguistic Society of America, Atlanta.

—— 2005*a*. Formal and empirical arguments for Morpheme Structure Constraints. Paper presented at the LSA Annual Meeting, Oakland.

—— 2005*b*. What L2 phonological effects tell us about the architecture of the phonological component. Paper presented at 14th World Congress of Applied Linguistics, University of Wisconsin-Madison, July 25, 2005.

—— 2006. What L2 phenomena reveal about phonological cognition. Paper presented at NAPhC 4, Concordia University, Montreal, May 13.

—— 2007. Phonological acquisition and theory construction. Paper presented at Oxford University, April 30.

—— and Nevins, Andrew. 2003. Underdetermination in language games: Survey and analysis of Pig Latin dialects. Paper presented at the Linguistic Society of America Annual Meeting, Atlanta.

Vaux, Bert and Nevins, Andrew. 2007. The shm-reduplication survey. Accessed online at <http://php-dev.imt.uwm.edu/prjs/markj/projects/fll_surveys/shm/index.php> on 2 June, 2007.

—— and Samuels, Bridget. 2004. Explaining vowel systems: dispersion vs. evolution. Paper presented at the Linguistic Society of America Annual Meeting, Boston.

—— —— 2005. Laryngeal markedness and aspiration. *Phonology*, 22.3: 395–436.

Vaysman, Olga. 2002. Against richness of the base: Evidence from Nganasan. *Berkeley Linguistic Society*, 28: 327–38.

Voegelin, Carl and Swadesh, Morrish. 1935. A problem in phonological alternation. *Language*, 15: 1–10.

Vroman, R. 1972. Problems in Old Norse phonology. Doctoral dissertation, University of Washington.

Wells, Rulon. 1949. Automatic alternations. *Language*, 25: 99–116.

Wheeler, Max. 2005. *The Phonology of Catalan*. Oxford: Oxford University Press.

Whitney, William. 1889. *Sanskrit Grammar*, 2nd edition. Cambridge, MA: Harvard University Press.

Wier, Thomas. 2004. Comparative markedness and opacity in Meskwaki palatalization. *Berkeley Linguistic Society*, 30.

Wiget, Wilhelm. 1916. *Die Laute der Toggenburger Mundarten*. Frauenfeld: Huber.

Wilbur, Ronnie. 1973. The phonology of reduplication. Doctoral dissertation, University of Illinois. Distributed by the Indiana University Linguistics Club, Bloomington, Indiana.

Wilson, Colin. 2000. Targeted constraints: An approach to contextual neutralization in Optimality Theory. Doctoral dissertation, Johns Hopkins University.

—— 2001. Consonant cluster neutralization and targeted constraints. *Phonology*, 18: 147–97.

—— 2003. Analyzing unbounded spreading with constraints: marks, targets, and derivations. Accessed on May 28, 2007 at <http://camba.ucsd.edu/courses/assimilation/readings/wilson-2003.pdf>.

—— 2006. Counterfeeding from the past. Manuscript. UCLA.

Wolfe, Andrew. 2000. Iterativity in Abkhaz stress. Manuscript. Harvard University.

Wolfram, Stephen. 2002. *A New Kind of Science*. Champaign, IL: Wolfram Media, Inc.

Yip, M. 1988. The obligatory contour principle and phonological rules: A loss of identity. *Linguistic Inquiry*, 19: 65–100.

Yip, M. 1996. Lexicon optimization in languages without alternations. In Jacques Durand and Bernard Laks (eds.), *Current Trends in Phonology: Models and Methods*, vol. 2. ESRI. Salford: University of Salford Publications, 757–88.

Zuraw, Kie. 2004. Class 19: Course summary through OT glasses. <http://www.linguistics.ucla.edu/people/zuraw/200A_2004/19OTII.pdf>.

Index of Authors

Numbers refer to pages, italicized numbers refer to footnotes on the corresponding pages.

This page intentionally left blank

Index of Subjects and Languages

Numbers refer to pages, italicized numbers refer to footnotes on the corresponding pages.