From:
Bridges between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman.
Edited by Donna Jo Napoli and Judy Ann Kegl.
Hillsdale NJ: Lawrence Erlbaum Associates. 1991.

13

Phonology as an Intelligent System

John Goldsmith
The University of Chicago

-Par où on commence? demanda Viale. Par le haut ou par le bas?
-Il n'y a pas de règles, observa Dumont.
Il retira ses lunettes, les astiqua.
-En général, par le haut, quand même, dit Dumont.
(*La mort dans une voiture solitaire*, p. 46 Hugues Pagan)

INTRODUCTION

The phrase "phonology as an intelligent system" suggests a contrast: a contrast with other views such as "phonology as an articulatory system," "phonology as a communicative system," "phonology as a social system," and "phonology as a mechanical system." Each of these views has something important to contribute to the study of phonology, but there is an important side of the matter that has been underplayed, and which today we should bring out and to the fore. The most interesting aspect of language is its role in the expression of human thought and intelligence, and yet until recently it seemed that there was a serious rift between those aspects of syntax and semantics that reflect thought, on the one hand, and the principles that govern phonology, on the other.[1]

[1]A recent perspective on this subject, but one taking a very different point of view, may be found in Bromberger and Halle, 1989. They suggested that phonology is fundamentally different from syntax in certain respects—which it indeed may be—but among the differences Bromberger and Halle suggested is the need for strict rule ordering in phonology. They offered one example, the well-known case involving the choice of the allophones of the diphthong [ay] in front of a voiced

This rift no longer gives the impression of being quite so immense and unbridgeable. This is not to say that phonology encodes propositional material; rather, the principles that govern the structure of the phonological components of a grammar, it is becoming clear, operate in accordance with more general principles that offer some hope of being understood within the larger context of cognition; and this is the possibility that I wish to consider. Thus we may emphasize here phonology as a cognitive system, one that organizes information first and foremost, one in which what is important is not the accidental outer form, the sound, associated with the elements of the phonological system, nor the social or communicative context, but rather the system of contrasts and constructs which is the essence of the phonological system within the grammar.

I focus on the goal-directedness of phonological processes in the following discussion, because there is a close connection between goal-directedness and intelligence. If we were to find a system that displayed no goal-directedness in its behavior, no matter how broadly construed, we would be hard-pressed to imagine a reason for calling the system intelligent. If, on the other hand, it did manifest some goal-directed behavior, then to that extent we would likely be willing to grant it a rudimentary portion of intelligence. Intelligence, for our purposes, we may take to be the ability to consider alternatives to being where one presently is, and to select the alternative that best suits one's current requirements. Phonological systems, in their own primitive way, I shall suggest, illustrate that kind of operation.

and a voiceless consonant in North American English: The diphthong in *write* is more central than it is in *ride*. The distinction between these two vowels is governed by rules, but is not lost when the consonant following (*t* or *d*) has been turned into a flap. They observed that *if* rules responsible for these processes are ordered, then the vowel-allophony rule must not be ordered after the flap-formation rule. This observation fails to make their point, though, for at least three reasons. First, it provides no argument that rules need to be linearly ordered; the two rules in question (ay-raising and flap-formation) could be unordered, applying simultaneously, and the correct result would result (see, for example, Lakoff [in press] for a long discussion of this point, or Kenstowicz & Kisseberth, 1979). Second, the context within which the rule of ay-raising applies is not, in fact, lost on the surface, that is, after flap-formation applies; again, crucial rule-ordering is not necessary, because there is a clear difference of phonetic vowel length in the syllable nuclei of the first syllable of *riding* and that of *writing*. From the point of view of Bromberger and Halle's argument, one could as well posit that length-difference is what determines the vowel quality of the diphthong. Third, as a development of the second point, the most important process involved in this area is not restricted to the diphthong *ay*, but holds more generally for all vowels, and involves the relative length of the vowel on the one hand, and the consonant following (*t, d*) on the other; we may say that the phonological feature of voicing is realized prominently in the determination of the ratio of the length of these two segments (vowel, consonant). How this calculation and realization is carried out will govern the distribution of the central versus noncentral allphones of *ay*. But this phonetic calculation is simply not the sort of process that is feasible within current phonological theory. The only representation for length within current phonological theory allows for integral units of length (1,2, perhaps 3; cf. Hayes, 1989, for example), and the differences at play in Bromberger and Halle's case are below these threshold differences; that is, the allophones of *ay* are both phonologically long, that is, associated with two moras. Hence current phonological theory would not even allow this rule to be a phonological rule, regardless of whether such rules could be ordered.

RULES IN CLASSICAL GENERATIVE PHONOLOGY

Phonological rules in classical generative phonology act, in each instance, as rules that modify a representation just in case their structural description is met (with further external conditions placed as well involving, for the most part, questions of rule ordering that we may comfortably leave aside for our present purposes). These rules' ability to effect a change in a representation comes, so to speak, from within; our conception of these rules is based on an implicit metaphor according to which these rules are internally powered—battery-operated, so to speak. Nothing further need be true for a rule to apply but that its structural description be satisfied. This conception of rules applying to representations is the generative inheritance from two sources: first, from logicians' formalization of logical derivation—in particular, Post's notion of a production system, and second, from historical linguistics' notion of regular sound change, in which ordered sequences of rules correspond simply and directly to stages in the evolution of a language.

I hope to show that this conception of rule application—which is by now thoroughly established in our modes of thinking—is both unnecessary and unsatisfactory, and that its rejection in no way entails a retreat or return to the static modes of thinking associated with structuralist conceptions. We can (and, as I will suggest, we have already begun to) establish a conception of phonology that largely (though not in every detail, to be sure) rejects this earlier governing metaphor, and replaces it with one that is more congenial to the modes of analyzing intelligence that have arisen in other disciplines.

CURRENT WORK IN PHONOLOGICAL THEORY

Work in autosegmental, metrical, and syllable phonology over the past 15 years has led us to a picture of phonology that is quite different in a number of ways from the image established in the classical period of generative phonology, the period influenced by *The Sound Pattern of English* (Chomsky & Halle, 1968). The most striking differences have been in the relative importance and articulation of the nature of phonological representations, on the one hand, and the class of phonological rules, on the other. In the classical period of generative phonology, representations consisted simply of linearly ordered strings of segments, themselves bundles of distinctive features. Today, complex multitiered structures are routinely explored to account more satisfactorily for phenomena from tone spreading to intrusive consonant insertion. In early generative phonolo-

gy, the syllable not only played no role, it had no way to be expressed; today it would be unthinkable to analyze a phonological system without something corresponding to the syllable, and both the internal and the external structure of the syllable are areas of ongoing research.

Phonological rules, in early generative grammar, were of considerable complexity, and problems of abbreviatory convention, of intrinsic and extrinsic ordering, and cyclic reapplication were of great importance. Now only the last, the problem of cyclic application, remains with us, and even it has been reformulated so as to help us come to grips with larger issues regarding the relationship between phonology and morphology.

In short, the balance of attention has shifted away from rules to problems of representation. Some have gone so far, in fact, as to deny the significance, or even the existence, of language-particular phonological rules. I shall explain some of my reasons for rejecting this later, but the tendency illustrates, by its extreme position, the shift that we are currently seeing in phonology.

Going hand in hand with the shift in emphasis towards problems of representation has been another shift which has by and large gone unnoticed up to now—or rather, it has been noticed only in bits and pieces, and the significance of the shift as a whole has not been apparent. With an articulated theory, or vocabulary, of phonological representations, it now becomes possible to make generalizations about phonological structures, and ask whether the phonological modifications that our phonological rules create are all pointing in a common direction or set of directions. Put simply, we may ask whether phonological rules uniformly modify phonological representations towards certain patterns, patterns at various different levels (using the term in a nontechnical sense for the moment): patterns regarding possible segments, possible syllables, possible feet, possible phonological words, and perhaps possible sequences of segments. To put it yet another way, we may ask whether there is not a sense in which phonological rules do more pulling (in particular directions) than pushing (away from the structural descriptions specified by a given rule); and whether even when they are pushing away from the structural description it is typically because of a more general property of the sound pattern of the language.

The answer to this question is, I believe, positive. Such an answer finds support in my own work, and draws together the work of many others currently working in phonological theory who have made less sweeping generalizations pointing in the same direction. Two brief examples might be helpful now, and we return to the matter in more detail later.

A growing (and by now overwhelming) body of literature on vowel epenthesis and deletion, beginning perhaps with Kisseberth's influential work (1970) on conspiracies, has established that the bulk of vowel epenthesis and deletion rules are sensitive to the syllable structure of the representation derived by the rule. A rule of epenthesis will typically apply just in case two conditions hold: its output contains sequences of well-formed syllables and its input is not properly syllabified—to put it simply, just in case its output is better than its input. To put the matter in such terms, of course, we need a general vocabulary and theory of syllabification, and as I have noted, we have taken many steps towards such an account in the last decade (for a recent discussion, see Itô, 1989). But the classical theory of generative phonology has no room at all for such notions; this theory is based on the notion of a rewrite rule that applies just in case its input conditions, or structural description, are met by a representation. A classical generative rule does not aim at any output or target structure; it is not, we may say, operating teleologically, with an eye to the structure that it is creating, and there is no sense in which we should understand it as aiming at a target schema. But that is just the property of vowel insertion and deletion rules that has emerged out of phonological research over the past two decades.

For example, Kisseberth (1970) pointed out that the epenthesis of the vowel *i* in Yawelmani Yokuts is the response of the phonological grammar to a situation where not all the phonological material is properly syllabified. Syllables in Yokuts may contain no more than one consonant in the onset and one in the coda, so sequences of three consonants can never be properly syllabified. In (1d), for example, the sequence of three consonants *gwh* is not syllabifiable as such, and the epenthetic vowel *i* is inserted in order to achieve proper syllabification of all of the phonological material. The hyphenation in the underlying and surface forms indicates breaks between morphemes; syllabification is not marked as such, but may be inferred from the generalizations just given.

(1)

surface	*underlying*	*surface*	*underlying*
a. xat-hin	/xat-hin/	xat-al	/xat-al/
b. bok'-hin	/bok'-hin/	bok'-ol	/bok'-al/
c. dos-hin	/do:s-hin/	do:s-ol	/do:s-al/
d. logiw-hin	/logw-hin/	logw-ol	/logw-al/

Similarly, early work in autosegmental phonology (Goldsmith, 1976) emphasized the importance of processes that spread autosegmental association over unbounded distance, up to (but not including) an already present association line. A good deal of controversy has attended the question of whether these automatic spreading processes can be uniformly universal, or whether they are to some extent language-particular. Regardless of the matter of universality, what is clear about such processes is that they are active processes aiming at a simple, particular target structure: one in which each vowel (for example) is associated with at least one tone, in the case of tone spreading, or one vowel harmony autosegment, in the case of vowel harmony, and so on. Spreading rules spread, in short, in order to create structures that are as saturated as possible—each vowel getting a tone, for example, when circumstances permit (and it is the rules that define whether the circumstances do in fact permit).

HARMONIC APPLICATION

The picture that emerges from examples like these, and many others, is one in which both target structures (or equivalently, phonotactics, or again, well-formedness conditions) and phonological rules play an important role, in a mutually supportive fashion, in a way that we may summarize as follows: All phonological rules apply in a harmonic fashion,[2] which is to say, they apply just in case their output is better than their input with respect to some criteria specified by a phonotactic (of the relevant level). In a word, then, rules apply for a good reason: in order to make a representation better match a pattern, or template, or phonotactic.[3] This is crudely put, to be sure; many of the most important operations involve patterns that are quite intricate, and other patterns involve structuration. For example, the single most important template towards which phonological rules move a representation is that according to which all segments are well integrated into a pattern of syllables. Thus, the erection of syllable structure, as well as of metrical structure, on a word is part of the pattern of a well-formed word that the phonological rules are pushing the representation toward. Patterns need not be merely at the level of overt sequences of phonetic segments; they may involve any item in the phonological vocabulary.

Such a notion smacks of the commonplace from the point of view of psychology, for example, where notions such as schemata—not to mention pattern recognition—are perfectly familiar. Such notions presuppose a global construction in which a number of properties are expected by the system to occur together. In the absence of reason to the contrary, a system utilizing schemata may use the information inherent to a given schema to increase the information available in a given situation, or even to modify information presently available. For example, believing that someone is a parent may lead us to further assume that they are adult, though that need not necessarily be true; and believing that someone has applied for a particular job and that he has not yet begun his dissertation may lead us to revise that second belief, on higher-order grounds: one would hardly be applying for such a job (we may reason) if one's dissertation were not done, or nearly done; we revise our assumptions in the light of our global knowledge. Phonological operations operate in certain parallel respects: Default specifications may be filled in, in accordance with both language-particular and universal principles, and phonological information may actually be changed on the basis of calculating the simplifications that would be achieved by modifying the representation in a derived environment (see the following discussion of lexical phonology for more on this).

[2]I allude here to Smolensky's harmony theory; see Smolensky, 1986.

[3]This notion has been discussed in similar contexts by Goldsmith (1989, in press a), Paradis (1988), Singh (1987, in press), Sommerstein (1974).

One thing that makes a system that understands special is that it shifts its representations in preestablished (or already definable) directions. That is, modifications of one's belief structures are made both in order to satisfy additional external information, of course, and in order to meet various internal conditions of coherence and simplicity: Defining and establishing such notions in formal and explicit ways is, to be sure, a difficult task, but to the extent that we succeed in doing it, what we expect of an intelligent system is that it should modify its representations in such a way that the structures better satisfy conditions of maximal coherence and simplicity.[4]

Each of these aspects of an intelligent system finds corresponding elements in phonology, I would suggest. The bulk of phonological rules apply in order to arrive at representations that maximally satisfy constraints (or, equivalently, schemata) that involve structuring phonological information.

If we may speak of harmonic application of phonological rules, we may also then consider speaking of a harmonic phonology, one in which this mode of rule application is central and essential, in ways that we will now clarify.

LEVELS IN HARMONIC PHONOLOGY

The picture that has emerged at this point may be described in the following way. A phonological description must include at least two things: a set of rules which describe the transitions that a given language permits, and a set of statements regarding relative well-formedness of various phonological structures. We may refer to this latter set of statements as phonotactics, and their role is to interact with the rules as described previously in relation to harmonic application: Rules apply just in case their output is better formed—better satisfies the phonotactics—than the input.

We may revise our mental image of this model in the following way. Rules specify permitted (and unordered) transitions between pairs of states (a word with and without a final consonant, for example, or with and without stress on the first syllable); these are language-particular statements, and can be conceived of as linking points on a large map that represents all possible phonological representations of a given language. The purpose of the phonotactics is to give a sense of peaks and valleys to that map, in such a way that the higher a representation is, the more poorly-formed (or less in step with the phonotac-

[4]This has nothing—or virtually nothing—to do with evaluation metrics of the sort considered and often discussed in generative grammar, which involve the issue of selecting a grammar on the basis of a given corpus of data—specifically, of selecting from a class of possible grammars which all satisfy the boundary conditions set by the observed data. On such a view of grammar selection (either as a methodology or as a theory of language acquisition), grammars are compared on the basis of simplicity; the matter discussed in the text involves the modification of representations within a grammar on the basis of simplicity considerations, broadly construed.

tics) it is, whereas the lower a representation is, the better it satisfies the phonotactics of the language. In such a picture, then, a representation will always seek the lowest position available to it through a sequence of permitted transitions on what we may call the *landscape* of that phonology.

Such an image is strikingly different from the image we have of traditional generative grammars, in part becaues of the much more lowly role played by rules in this picture, which goes so far as to virtually suggest that rules can be conceived of as being replaced by representations, though in this case the representation is not of any particular form, but rather is a representation of the sound pattern of the entire language.

This picture is useful for some aspects of phonological analysis, and not useful for others. It is especially useful for understanding those aspects of phonological analysis that involve considerable feeding orders, for example, and in which there are no counterfeeding orders. Syllabification, for example, typically involves an interaction of a large number of processes, such as coda formation, onset formation, epenthesis, vowel deletion, and foot formation (i.e., stress assignment); a similar observation holds for the process of foot formation as well. In most cases, these processes can be reformulated as involving constraints whose simultaneous solution represents the correct or observed pattern.[5,6]

However, wha makes phonology strikingly different from other aspects of grammatical theory is that one simply cannot establish a set of phonotactics, or constraints, for the phonology of a given language, and leave it at that; put in a more traditional way, there are significant rule interactions not of the sort just mentioned, and there are rules whose effects are not harmonic.

Consider, for example, some well-known facts from Lardil, an Australian language analyzed by Hale (1973), and discussed as well by Kenstowicz and Kisseberth (1979). In Lardil, a word may end with an open syllable (i.e., with a vowel), as in (2a), or with an apical consonant (2b), but with no other consonant. If a word ends underlyingly with one or more nonapical consonants, the consonant(s) are deleted, so as to satisfy the condition on how the Lardil word may end. This is illustrated by the examples in (2c), in which a stem

[5]The insufficiency of Lamb's (1962 and elsewhere) stratificational models was argued by Postal (1968) on the grounds of the commonness of feeding orders in natural languages. A revised and updated version of this argument was made by Lachter and Bever (1988, pp. 201-03) against connectionist views of phonology. These arguments do not transfer to the present framework, where feeding orders, or their equivalent (cf. footnote 5), are permitted; cf. also footnote 1.

[6]From a more radical perspective, adumbrated in the section on cyclicity below and discussed in more detail in Goldsmith(in press b), the effects of harmonic rule application result not from an algorithmic procedure that applies discrete rules in a linear sequence, but rather would result from the phonological model itself being implemented in a network where the presence or absence of a given feature or structural relation can be assigned a real number with a value from -1 to +1. In such a scheme, the rules of the level may be replaced by statements of local connections between neighboring elements, and the "output" of the level is the equilibrium value that the network settles into. This approach is developed in Goldsmith (in press b) for some cases involving the metrical grid and stress patterns.

such as *ngalu* (underlyingly *ngaluk*) loses its final consonant in the uninflected form, surfacing then as *ngalu*. (*th* represents a single, laminal dental consonant, and *ng* represents a velar nasal throughout, *ṭ* is an apicodomal stop, and *r* is apicodomal as well.)

(2) a.

tipiti	species of rock cod
mela	sea
wanka	arm
kungka	groin
nguka	water
kaṭa	child
ngawa	wife
ngalu	story
putu	short
murkuni	nullah
ngawunga	termite

b.

yalul	flame
mayar	rainbow
wiwal	bush mango
karikar	butter-fish
yiliyil	species of oyster
yukar	husband
wulun	species of fruit
wuṭal	meat
kantukan	red
karwakar	species of wattle

c.

underlying form of stem	*uninflected*	*non-future*	*future*	*gloss*
thurarang	thurara	thurarang-in	thurarang-kur	shark
ngaluk	ngalu	ngaluk-in	ngaluk-ur	story

If a word with three or more syllables ends in a vowel underlyingly, however, it loses that vowel, as in the uninflected (first column) forms in (3a), and if this vowel loss leads to a situation in which a nonapical consonant is now[7] word-final, then these nonapical consonants are lost, just as before; this is illustrated in (3b). The crucial point for us is that the loss of the vowel is not motivated by a need to satisfy a phonotactic, for word-final vowels are perfectly satisfactory, and the shift from (e.g.) *munkumunku* to *munkumunk* (which is then followed by the deletion of the word-final consonants; cf. footnote 4) is one that

[7]We slip here into derivational idiom, but only for a moment, and it should be taken as a touch of irony.

moves away from satisfaction of the phonotactics. Hence there are at least some rules with this nonharmonic property, and we must deal with that fact. (*tj* represents a single, laminal alveopalatal consonant, and *th* a single laminal dental consonant).

(3) a.

stem	*uninflected*	*nonfuture*	*future*	*Gloss*
yalulu	yalul	yalulu-n	yalulu-r	flame
mayara	mayar	mayara-n	mayara-r	rainbow
wiwala	wiwal	wiwala-n	wiwala-r	bush mango
karikari	karikar	karikari-n	karikari-wur	butter-fish
yiliyili	yiliyil	yiliyili-n	yiliyili-wur	species of oyster

b.

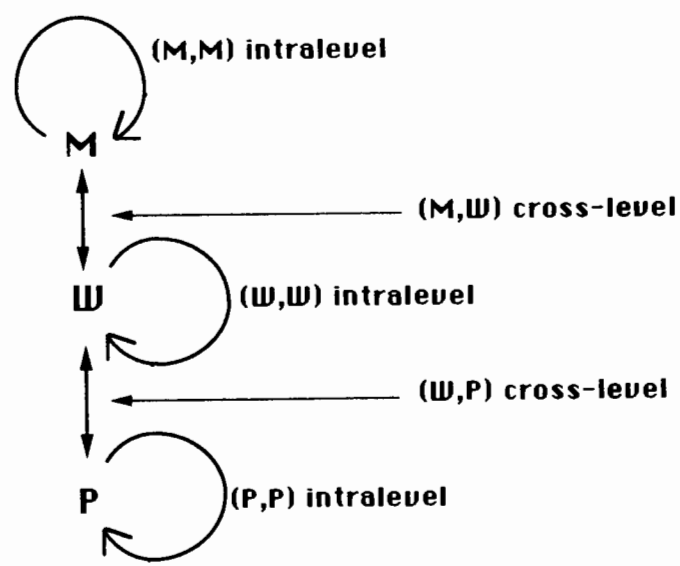
putuka	putu	putuka-n	putuka-r	short
murkunima	murkuni	murkunima-n	murkunima-r	nullah
ngawungawu	ngawunga	ngawungawu-n	ngawungawu-r	termite
tipitipi	tipiti	tipitipi-n	tipitipi-wur	species of rock-cod
thaputji	thapu	thaputji-n	thaputji-wur	older brother
munkumunku	munkumu	munkumunku-n	munkumunku-r	wooden axe

These hard, unpleasant facts of phonological life force us to recognize that the image of rules as transitions on a phonological landscape is only a part of a larger picture, and that that part corresponds to the traditional notion of levels in linguistic theory. That is, what constitutes a level, in traditional terms, is a set of generalizations regarding the linguistic representation; these generalizations may be restated in their entirety as phonotactics (or, in our metaphor, as statements regarding what is higher than what, and what is lower than what, on the landscape). A level is not, then, one stage in a derivation; it is not even a single representation: It is (and this is the point of this paper) a set of phonotactics, and a representation of a given utterance U on a level L is a path from a starting point R_1 to a final resting point R_n. The final resting point R_n is that representation which is the best-formed (i.e., the lowest on the landscape) of all points accessible to R_1 via the paths made available by the rules of the language on that level.

Each level, then, contains a set of rules, which we may refer to as intralevel rules, and these rules will necessarily apply in a harmonic fashion. But as we have just suggested, there is more than one level in a phonology of a natural language. A fair amount of exploration suggests that although two phonological levels is inadequate, a model with three levels is sufficiently rich to deal

with the phenomena that have come to light. Such a model will contain three levels—which we will refer to as M-level (essentially the underlying, or morphophonemic, level), the W-level (the level at which pure syllable structure is established), and P-level (phonetic level). Each level consists of the statement of tactics at its level, plus a set of intralevel rules. The three phonological levels, we may assume, relate to each other in much the same way that the other linguistic levels relate to one another (I draw heavily here on Sadock 1985, 1990). That is, the relation between the W-level and the M-level is logically parallel to that between the syntax and the morphology: the two tend to line up, in general, in a natural way, but do not need to do so in any particular case. The rules that relate levels (whether they be M-level or syntactic) are interface rules, in Sadock's terminology (or cross-level rules). In principle, then, there should be six classes of phonological rules: three intralevel rule sets—M-level, W-level, and P-level intralevel rules: (M,M); (W,W); (P,P), and three cross-level rule sets: (M,W); (W,P); and (M,P). If there is a hierarchization of levels in phonology, the last one—(M,P)—may not exist (as I shall assume for expository reasons), and we would arrive at a picture as follows:

(4)



M/W/P model

FIG. 13.1.

Ultimately, the three levels of phonological theory should be viewed as not different in kind from the other levels of grammatical theory, such as the morpho-

logical[8], the (two) syntactic levels, an argument-structure level, and so forth.

This perspective requires us to consider all nonharmonic rule effects as cross-level (or interface) rules; this has as a consequence that there can be no more than two such rule applications in any given phonological derivation. The rule deleting a word-final vowel in Lardil is thus a cross-level (M,W) rule, and the rule that eliminates illicit word-final consonants (i.e., nonapical word-final consonants) operates as a (W,W) rule.

Although rules applying within a level give rise to what appear to be quite transparent rule interactions, the system as a whole need not have that property, no more than any familiar generative system. The constraints and the rules on the W-level and the P-level may be sufficiently different that the effect of having both levels (with their rules) is one of a reasonable degree of complexity. Consider the case of Yup'ik, for example, as discussed by Jacobson (1985).[9] Although there is a good deal of variation among the various dialects of Yup'ik in Siberia and Alaska, in Central Alaskan Yup'ik, stress is assigned to all bimoraic syllables, as well as to certain other syllables: to word-initial closed syllables, and to every other syllable (the even-numbered syllable) in sequences of light syllables (excluding the case of closed plus open light syllables), as illustrated in (5), where italicized syllables are stressed; word-final syllables are never stressed.

(5) a.	*aang* qagh *llagh* llang *yug* tuq	"he wants to make a big ball"
b.	ang *yagh* llagh *llang* yug tuq	"he wants to make a big boat"
c.	qa *ya* ni	"his own kayak"
d.	qa *yaa* ni	"in his (another's) kayak"
e.	sa *qu yaa* ni	"in his (another's) drum"
f.	qa *ya* pig *ka* ni	"his own future authentic kayak"
g.	qa *ya* pig *kaa* ni	"in his (another's) future authentic kayak"
h.	a *te* pik	"real name"
i.	ang *yagh* lla ka	"my big boat"
j.	ang *yagh* lla kaa	"it is his big boat"

[8]There is occasionally a confusion between the morphological level and the M-level. The M-level consists of elements that are essentially phonological: The utterance *the dog is asleep* consists of twelve phonological segments on that level. On the morphological level of analysis, this expression has five units present, including the copula, the Present tense morpheme, and the single, atomic morpheme *dog*.

[9]The examples given here are from Jacobson (1985), and are discussed as well in Goldsmith (1990), where I unfortunately failed to cite Jacobson directly, giving only the name of the volume in which his work appears (Krauss, 1985). This example is discussed in the text in essentially the terms used by Jeff Leer, to whom I am indebted, in unpublished work. In addition to the Jacobson paper and others in Krauss (1985), see especially Leer (1985). Anyone who has looked at the prosodic systems of Yup'ik and related languages will know full well that any single, simple statement risks being an unfortunate oversimplification; I trust I have not oversimplified to the point of inaccuracy, and Leer is not to be taken as responsible for any oversimplifications that I have brought into the picture.

These patterns are established at the W-level, and involve reference to two types of conditions at this level. First, and quite generally across languages, there is a preference for bimoraic syllables to be stressed, rather than unstressed (a condition referred to by Prince, 1983, as Quantity-Sensitivity). Stress is assigned to satisfy this requirement. In addition, syllables must be organized into feet, and these are iambic in Yupik (i.e., weak-strong).[10] In short, at this level, stress is assigned to match inherent quantity of the syllables. At P-level, a rather different process occurs, by which syllable weight is modified to match the stress pattern that was established at the other level. In essence, what happens is that if a syllable is stressed, it must be heavy; if it is already—inherently—heavy by virtue of having a long vowel or being closed, that is sufficient; otherwise, the syllable is made heavy by one means or another (essentially, lengthening the vowel of the syllable unless that vowel is a schwa, in which case the consonant of the following onset is geminated, creating a

[10]Although there is a relationship between these two principles, it is not one that needs to be conceived of in terms of derivational rule ordering. What is crucial is that the grammar capture the fact that the first generalization—Quantity-Sensitivity—is a stricter generalization than the second (that syllables are organized into iambic (weak-strong) groupings). That is, if a grouping of syllables into weak-strong/weak-strong/. . . should attempt to put a long-voweled syllable in a weak position of a foot, an inappropriate structure would result. This can be conceived of in several ways, one of which is ordering quantity-sensitivity first (others include: making the algorithm that assigns iambic feet directly relevant to the inherent quantity of the first part of the foot; or treating the long voweled syllable as composed of two units over which an iambic foot must be erected without going any further. I ignore these various possibilities for expositional reasons; see Goldsmith (1990, chapter 4), and Krauss (1985) for further discussion). But nonderivational conceptions of this relationship are possible as well. It goes beyond the scope of this paper to go into the question in the detail it deserves, but we may address it informally here. As suggested in the text, we must consider that the addition of metrical structure (here, at W-level) to a series of syllables that does not have metrical structure is a "descent" on the energy landscape, that is, a decrease in total complexity. If there were no long syllables in a word to worry about, then from the point of view of an energy landscape, the establishment of binary feet throughout would be analyzed as the result of our assigning a certain cost C_1 (i.e., a height in energy space) to a syllable *not* being part of a foot, and also of assigning a cost C_2 to the establishment of each foot. Thus it is better for each syllable to be in a foot than not to be in one; but feet do not come free. Only if C_2 is greater than zero can we be sure that we will not simply assign foothood (i.e., stress) to each and every syllable: There must be a "cost" assigned to setting up such feet. (Clearly, C_2 must be less than two times C_1 as well—that is, the cost of setting up the foot must be no greater than the reward we get for act, which is a savings of amount C_1, twice over, once for each syllable that is placed in the foot.) Now, to ensure that Quantity-Sensitivity as a generalization has precedence over the assignment of iambic feet, in a framework without derivations and the ordering that they assume, it is sufficient to consider the difference D in energy height between that assigned to a long-vowel syllable that is the head of its foot and that assigned to a long-vowel syllable that is not the head of its foot. If we establish D as being greater than C_2 (the "cost" of establishing a new foot), then we will get the desired result. Put another way, the only thing that blocks setting up a new foot for each syllable is that feet "cost" something (C_2, by definition); but if we make sure that the profit derived from assigning a foot to a heavy syllable is positive (i.e., the proceeds exceed the necessary costs [$D > C_2$]), then we will get the desired result.

closed syllable, except in Central Siberian). This is illustrated in (6). Thus the effect of the generalizations on the two distinct levels is to make the effects of each level less than obvious, even though the effects within each individual level are simple and direct. Each level—W and P—strives to achieve a simple matching between the accent and the weight of the syllables, though the two levels achieve this (to the extent that they succeed) in opposite directions.

(6) a. W-level: qa *ya* pig *ka* ni "his own future authentic
surface: [qa *ya:* pix *ka:* ni] kayak"
b. W-level: ang *yagh* lla ka "my big boat"
surface: [ang *yaɣ* la ka]

LEXICAL PHONOLOGY

Lexical phonology (Kiparksy, 1982) makes a particular suggestion that has not been especially pursued by most phonologists endorsing that research program, one that is relevant to our discussion (see also Goldsmith, 1990, chapter 5). The suggestion is that the class of lexical phonological rules is coextensive with the set of rules that establishes markedness for lexical entries. For example, if the rule of trisyllabic shortening (7) is a lexical phonological of English, operative in such hoary examples as *divine/divinity* to shorten the first vowel of the suffixed, derived nominal, it also functions to express the generalization that any vowel followed by an unstressed syllable and another syllable ought to be short, and will be long only under marked conditions; in that sense, *Canada*, with its short first vowel, is better than *rudiment*, with its long first vowel.

(7) V → [-long] / — C_0 [V,-stress] C_0 V

Lexical phonology unfortunately offers no explicit means for the language learner to figure out what the lexical redundancy rules of his or her language are, but it does suggest that once such rules have been established, they are now operative in analyzing morphophonemic alternations, or in lexical phonology's terminology, they function as lexical phonological rules. Put another way, if we take statements of markedness with regard to lexical redundancy to be contributions to the statements of relative wellformedness on either M-level or W-level, with better formed (i.e., less marked) representations being lower on the landscape then minimally different, but marked, representations, then lexical phonological rules will always make a representation move downhill, that is, harmonically. Repeating the last example, if a short vowel if less marked than a long vowel in the position / — c [v, -stress] c v, then when a long vowel becomes short (in *divin-ity*) during the derivational process that is responsible for the deadjectival nominal, the shift involved is one that simplifies the

representation, or pushes the representation downhill. Thus even the rules of lexical phonology, understood in this way, have the harmonic property that we are focusing upon.

CYCLICITY

The simple model described earlier in (4), Fig. 13.1, with its three levels, appears to say nothing about the concept of cyclicity, a notion central to lexical phonology and a good deal of recent work in phonology. The present model does offer an interesting and attractive reanalysis of some of the fundamental properties of a cyclic account however.

On most accounts, the notion of cyclicity involves particular details of rule application and reapplication. For lexical phonology, which is heavily committed to a processual and derivational conception of phonological analysis, cyclic strata are organized in such a fashion that after each successive affix is attached, a sequence of phonological rules is applied, as their individual structural descriptions are met; there will be as many opportunities for the entire set of rules to apply as there are affixes attached.

This notion of cyclicity has no place in the present model, because the overwhelmingly derivational model that is assumed by lexical phonology has no place here. Let us take the opportunity to step back and observe what is involved in considerations of cyclicity. We find in general two schools of thought of the subject. On the one hand, there is the word-based school of cyclicity, discussed in Brame (1972a, 1974), Aronoff (1976), Harris (1983), Kiparksy (1982), and Goldsmith (1990), according to which the word is the unit to which further operations may be performed to yield derived words: schematically, as in (8); the domains marked "W[ord]", and no units smaller, are subject to cyclic reapplication. On the other hand, there is another view of cyclicity according to which cyclicity has nothing to do with the phonological word, but devolves rather from the dynamic process of word formation, as discussed in (for example) Chomsky and Halle (1968), and more recently, Poser (1989).[11] The last example is useful in establishing a contrast between these conceptions.

(8) [w [w a] b];

Poser (1989), based on work of Peter Austin, discussed the case of Diyari, a language in which stress is assigned to alternate syllables, starting on the left, within each morpheme of a word, as illustrated in (9), with morpheme-final syllables not receiving stress in any event. Rather than allow a grammar

[11]Poser cited lexical phonology throughout, seemingly unaware that Kiparsky's (1982) statement of lexical phonology requires that cyclic domains be minimally words.

the ability to say such a thing directly, Poser suggested that the effect should be derived indirectly, in the following way: A cyclic analysis, as he described it, will leave visible only the root on the first cycle, and on each successive cycle the grammar will find one more affix than it did on the cycle before. Therefore, he suggested, a cyclic account may assign alternating (left to right) stress to the root, and again on each successive cycle, just so long as the stress assignment done on an earlier cycle is left untouched on the later cycle(s). In such a way, each cycle will affect only the material that is new on that cycle, and by the way things have been set up, each affix will have exactly one cycle during which it is the new affix to which no stress has yet been assigned. Crucial to Poser's account is that the morphemes in question are in no way and in no sense words.

(9) a. ŋádawàlka-tâdi "to close + passive"
b. yákalka-yìrpa-màli-na "ask-benefactive-recipient-partative"

Such a view of cyclicity has little or nothing in its favor in this case (or others), as far as I can see, except that it permits one an indirect fashion of saying what one might as well say directly, which is that the relationship established between syllables and the metrical grid may be sensitive to morphemic identity—just as tone-syllable initial association may be, in many tone language (e.g., in Llogoori; Leung, 1986; Goldsmith, in press c), where the initial tone association of each tone must be to the leftmost vowel of the morpheme that is logically associated with that tone; the process is thus a morpheme-by-morpheme process, not a word-level process.[12]

The real significance of cyclicity, as Brame and the others cited earlier argued, is that there are phonological cases in which one can argue that there is a nested bracketing of phonological words, as in $[_{W\text{-}1}\ [_{W\text{-}2}\ a\ ]\ b\ ]$; cf. (8). Cyclicity then enters into the analysis in two ways: first, phonological processes may be effected within W_2 because it is a phonological word, processes that would not occur otherwise (i.e., processes that would not occur if the material marked as "a" were not treated as a word); and second, effects that we otherwise expect to take place within a phonological word may be blocked across the boundary separating W-1 and W-2, that is, between the base and the suffix. The first case is exemplified in Selayarese (as discussed by Mithun & Basri, 1986),where we find that stems that end in *s, l,* or *r* must have an epenthetic vowel (identical to the preceding vowel) added to them if they are to serve as full phonological words; /lamber/ "long" therefore surfaces as [lambere], for example. When a suffix such as *-ang* "comparative" is added, the base is not treated as a separate word, and we find no epenthetic vowel, as in /lamberang/ "longer." However, there are other suffixes which attach to units that must be analyzed as full phonological words; the first person possessive suffix *ku*, for example, attaches to the word *sahala* "profit" (from underlying /sahal/), to give the complex form /sahalakku/, which has the structure $[_W\ [_W$ sahala] ku].

[12]This point was first made, I believe, in Clements (1983). In general, cyclicity will not help in cases of the following sort: where on a "new cycle," both suffixal material on the skeletal tier and suffixal material on the tonal tier is added, and where the final vowel of the base (i.e., the material already present on the previous cycle) was not associated with a tone. If the tonal material associates with the suffixal skeletal material, rather than with the leftmost available vowel in the base, as is the case, for example, in Llogoori, then it is simply necessary to allow tone-to-skeleton association to be directly sensitive to morphemic identity—the very possibility that Poser's discussion presumes should not be allowed.

Concerning the second effect of cyclicity—or, as we may equally refer to it, recursive phonological word-structure—we find cases as in, for example, English *Indianaism* (a speech pattern peculiar to Indiana) in which phonotactic regularities that otherwise hold for English are blocked across a boundary. Here, we have a sequence of schwa plus high vowel, which can otherwise be found in English only across full word boundaries, but never inside a single phonological word. When the suffix *-ism* is attached to a base without the recursive word structure of (8), the schwa is deleted (as in *buddha + ism > buddhism*); but this process of deletion does not happen to the schwa at the end of the inner cycle in [[indiana]ism], an example of the second type of cyclicity effect.[13]

These two effects are, I believe, the only robust effects that can be attributed to cyclicity, and both can be reconstructed from a point of view that reconstructs derivations in the way I have suggested in this paper.[14] Regarding the first point, if a subpart of a larger phonological word is itself a phonological word, as in (8), then it must satisfy the language's tactics for being a well-formed word, just as an embedded clause must satisfy all the grammatical conditions for being a clause, even though it may well be (irrelevantly) embedded within a larger, matrix clause. Regarding the second point, we must observe that it is still an open question as to which word-level phonological rules are blocked from applying across word-boundary (so to speak) as in a structure like (8). The simplest account would be one according to which no word-level rules apply strictly across such boundaries; in those cases where rules appear to, one of two alternatives may be the case: (a) in the case of rules such as

[13]A similar case can be found in Hall, 1989, where the distribution of German [ç] and [x] is explored from the point of view of lexical phonology. As Hall pointed out, [x] appears after a back vowel, and [ç] essentially elsewhere, but this generalization must be restricted to take the phonological word as an absolute barrier, as seen in a form such as the classic *Kuhchen* "cow (diminutive)," which has the form [[ku] çən]. As Hall observed, attempts to formulate this observation in derivational terms consistent with the principles of lexical phonology regarding the interleaving of phonological and morphology leads ineluctably to violations of other principles that are equally central to lexical phonology.

[14]From one point of view, this should hardly be surprising: The two cyclicity effects that I have reviewed in the text have precise analogs in syntax, and the claim has been established that the syntactic cycle can be reinterpreted in (or rather, reduced to) nonderivational terms.

Trisyllabic Shortening, applying to *divin-ity* to form *divinity* with a short second vowel, the phonological structure is not as in (8), but simply [divinity]: that is, phonological structure need not match morphological structure (or, to put it another way, word-based morphology need not always give rise to nested phonological word-structure); (b) in the case of stress rules, as Halle and Vergnaud (1988) demonstrated, each word-cycle may construct its own metrical grid, independent of the grid associated with the embedded phonological word; this gives the appearance of the grid constructed with outer word cycle overriding that constructed on an embedded cycle.

DISCUSSION AND CONCLUSION

The picture that emerges of the phonological system, then, is one in which rules serve as a means for getting representations to maximally satisfy phonotactics of the individual phonological levels of the grammar. How, we may ask, does this picture fit in with other conceptions of grammar and of cognition?

Recent work on connectionism speaks in a kindred fashion. Rumelhart and McClelland, for example, offered the following observation,

> Imagine a computational system that has as a primitive, "Relax into a state that represents an optimal global interpretation of the current input." This would be, of course, an extremely powerful place to begin building up a theory of higher level computations... These sort of primitives... are the kind of emergent properties that PDP mechanisms give us, and it seems very likely that the availability of such primitives will change the shape of higher level theory considerably. (Rumelhart & McClelland, 1986, pp. 126-127)

This appears to be exactly the sort of higher-level vocabulary that is required by the type of phonology—harmonic phonology—that I have adumbrated in this paper. Various discussions in the current literature have raised questions regarding the relevance of connectionist modeling to linguistic problems (for example, Lachter & Bever, 1988; Pinker & Prince, 1988). I interpret the difference between their pessimism and my optimism as based largely on how satisfied one is that the current models of phonology (or grammar, more generally) are within shouting distance of the final truth. If our current derivational models are—minor details aside—essentially correct models of the truth, then connectionist revisions are neither welcome nor helpful. If, on the other hand, serious reconsideration of even the most basic questions of the organization of phonological derivations and rule application are the order of the day, as I have suggested here, then it is certainly within the realm of the conceivable that the types of generalizations that emerge from connectionist models may be closer to the sort that we need in the newer model of phonology.[15]

[15] I have made some concrete proposals along these lines for the treatment of stress in Goldsmith, in press c, and with Gary Larson, for syllabification in Goldsmith and Larson, 1990. See also Larson, 1990.

I have mentioned several possibilities in this paper that concern what comes close to being a nonderivational phonology. The possibilities of a nonderivational syntax have been discussed and explored considerably over the past decade or more; few serious candidates for anything parallel have arisen in phonology, precisely because a static conception seems so unappealing in the face of all that we know about phonological systems in natural languages. What I have suggested in this paper amounts to a proposal to factor the dynamic character of phonological analyses into a number of subsections, corresponding to individual linguistic levels, in such a way that we can identify the phonological dynamic in each case as an instance of maximally satisfying the constraints of that particular level. If this program can be satisfactorily extended to the whole of phonology (and then, presumably, grammar as a whole), we may well find ourselves in a position in which our linguistic model satisfies simultaneously the requirements of a psychologically real model and those of a linguistically complete model.

ACKNOWLEDGMENTS

I am very happy to offer this paper to this volume for Lila Gleitman, my first linguistics teacher. She has often insisted on the importance of bringing together considerations of linguistic evidence and of psychological reality, and it may go without saying that the considerations discussed in this paper arose largely out of the psychological implausibility of current generative accounts of phonology. This is an abbreviated version of a longer work (Goldsmith, in press) to be published in a collection edited by myself, and also reflects some suggestions made in Goldsmith 1989a and 1990. I am grateful to many people for discussions or suggestions that led to the conclusions here, including Anna Bosch, Diane Brentari, Morris Halle, Gary Larson, Jeff Leer, John McCarthy, K.P. Mohanan, Carole Paradis, Jerry Sadock, Ivan Sag, Raj Singh, and Caroline Wiltshire.

REFERENCES

Aronoff, M. (1976). *Word-formation in generative grammar*. Linguistic Inquiry Monograph Series 1. Cambridge, MA: MIT Press.

Brame, M. (1972a). The segmental cycle. In M. Brame (Ed.), *Contributions to generative phonology*. Austin: University of Texas Press.

Brame, M. (Ed.). (1972b). *Contributions to generative phonology*. Austin: University of Texas Press.

Brame, M. (1974). The cycle in phonology: Stress in Palestinian, Maltese, and Spanish. *Linguistic Inquiry, 5*, 39-60.

Bromberger, S. & Halle, M. (1989). Why phonology is different. *Linguistic Inquiry, 20*, 51-70.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Clements, G. N. (1983). *Some parameters of variation in tone languages*. Paper presented at the Conference on Hierarchy and Constituency in Phonology, University of Massachusetts at Amherst.

Goldsmith, J. (1976). *Autosegmental phonology*. PhD dissertation, MIT.

Goldsmith, J. (1989). Licensing, inalterability, and harmonic rule application. In R. Graczyk, B. Music, & C. Wiltshire (Eds.), *Papers from the 25th Annual Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

Goldsmith, J. (1990). *Autosegmental and metrical phonology*. Oxford, England and Cambridge, MA: Basil Blackwell.

Goldsmith, J. (in press a). Harmonic phonology. In J. Goldsmith (Ed.), *The last phonological rule: Reflections on constraints and derivations*.

Goldsmith, J. (in press b). Local modeling in phonology. In S. Davis, (Ed.), *Connectionism: Theory and practice*. Vancouver: University of British Columbia Press.

Goldsmith, J. (in press c). Tone and accent in Llogoori. In D. Brentari, G. Larson, & L. Macleod (Eds.), *The joy of syntax: Papers in honor of James McCawley*. Amsterdam: Benjamins.

Goldsmith, J. and Larson, G. (1990). Local modeling and syllabification. In K. Deaton, M. Naske, & M. Ziolkowski, *Papers from the 26th Annual Regional Meeting of the Chicago Linguistic Society, Part Two: Parasession on the Syllable in Phonetics and Phonology*.

Hale, K. (1973). Deep and surface canonical disparities in relation to analysis and change: An Australian example. *Current Trends in Linguistics* 11, 401-458.

Hall, T. A. (1989). Lexical Phonology and the distribution of German [ç] and [x]. *Phonology*, 6(1): 1-17.

Halle, M. & Vergnaud, J. R. (1988). *An essay on stress*. Cambridge, MA: MIT press.

Harris, J. W. (1983). *Spanish syllable structure and stress: A nonlinear analysis*. Linguistic Inquiry Monograph 8. Cambridge, MA: MIT Press.

Hayes, B. (1989). Compensatory lengthening in moraic phonology. *Linguistic Inquiry, 20*, 253-306.

Itô, J. (1989). Prosodic theory of epenthesis. *Natural Language and Linguistic Theory*, 7, 217-260.

Jacobson, S.A. (1985). Siberian Yupik and Central Yupik Prosody. In M. Krauss (Ed.), *Yupik eskimo prosodic systems: Descriptive and comparative studies*.

Kenstowicz, M. & Kisseberth, C. (1979). *Generative phonology*. New York: Academic Press.

Kiparsky, P. (1982). Lexical morphology and phonology. In I.-S. Yang (Ed.), *Linguistics in the morning calm* (pp. 3-91). Seoul: Hanshin.

Kisseberth, C. (1970). Vowel elision in Tonkawa and derivational constraints. In J. M. Sadock & A.L. Vanek (Eds.), *Studies presented to Robert B. Lees by his students* (pp. 109-137). Champaign: Linguistic Research Inc.

Krauss, M. (Ed.). (1985). *Yupik eskimo prosodic systems: Descriptive and comparative studies*. Alaska Native Language Center Research Papers Number 7. Fairbanks: Alaska Native Language Center.

Lachter, J. & Bever, T.G. (1988). The relation between linguistic structure and associative theories of language learning: A constructive critique of some connectionist learning models. In S. Pinker & J. Mehler (Eds.), *Connections and symbols*. Cambridge, MA: MIT Press.

Lakoff, G. (in press). Cognitive phonology. In J. Goldsmith (Ed.), *The last phonological rule*.

Lamb, S. (1962). *Outline of stratificational grammar*. Berkeley: University of California.

Larson, G. (1990). Local computation networks and the distribution segments in the Spanish syllable. In K. Deaton, M. Noske, & M. Ziolkowski, *Papers from the 26th Annual Regional Meeting of the Chicago Linguistic Society, Part Two: Parasession on the Syllable in Phonetics and Phonology*.

Leer, J. (1985). Toward a metrical interpretation of Yupik prosody. In M. Krauss (Ed.), *Yupik eskimo prosodic systems: Descriptive and comparative studies*.

Leung, E. (1986). *The tonal phonology of Llogoori: A study of Llogoori verbs*. Unpublished master's thesis; Cornell University.

Mithun, M. & Basri, H. (1986). The phonology of Selayarese. *Oceanic Linguistics, 25*, 210-254.

Paradis, C. (1988). On constraints and repair strategies. *The Linguistic Review, 6*, 71-97.

Pinker, S. & Prince, A. S. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. In S. Pinker & J. Mehler (Eds.), *Connections and symbols*. Cambridge, MA: MIT Press.

Poser, W. (1989). The metrical foot in Diyari. *Phonology, 6*, 117-148.

Prince, A.S. (1983). Relating to the Grid. *Linguistic Inquiry, 14*, 19-100.

Rumelhart, D. & McClelland, J. (1986). PDP models and general issues in cognitive science. In D.E. Rumelhart, J.L. McClelland, & the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Sadock, J. (1985). Autolexical syntax: A proposal for the treatment of noun incorporation and similar phenomena. *Natural Language and Linguistic Theory, 3*, 379-440.

Sadock, J. (1990). *Autolexical syntax*. Chicago: University of Chicago Press.

Singh, R. (1987). Well-formedness conditions and phonological theory. In Dressler et al 1987.

Singh, R. (in press). On repair strategies and constraints: A reply to Paradis. *Linguistic Review, 6*.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. In D.E. Rumelhart, J.L. McClelland, & the PDP Research Group. *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Sommerstein, A.H. (1974). On phonotactically motivated rules. *Journal of Linguistics, 10*, 71-94.