Learning Long-Distance Phonotactics

Jeffrey Heinz

This article shows that specific properties of long-distance phonotactic patterns derived from consonantal harmony patterns (Hansson 2001, Rose and Walker 2004) follow from a learner that generalizes only on the basis of the order of sounds, not the distance between them. The proposed learner is simple, efficient, and provably correct, and does not require an a priori notion of tier or projection (contra the model in Hayes and Wilson 2008); nor does it rely on the additional structure provided by Optimality Theory grammars (Prince and Smolensky 1993, 2004) or grammars in the principles-and-parameters framework (Chomsky 1981, Dresher and Kaye 1990, Gibson and Wexler 1994). Not only does the noncounting nature of nonlocal dependencies automatically follow from the way the learner generalizes, it also explains the absence of blocking patterns from the typology. Finally, the learner lends support to the idea that long-distance phonotactic patterns are phenomenologically distinct from spreading patterns, contra the hypothesis of Strict Locality (Gafos 1999, et seq.).

Keywords: phonotactic learning, long-distance agreement, precedence, grammatical inference, formal language theory

1 Introduction

1.1 The Contribution

People learn language, but it remains a mystery how they, or any other computing device, can do so. The perspective provided by formal learning theory (Vapnik 1998, Jain et al. 1999) and the related discipline of grammatical inference (de la Higuera 2010) aligns closely with linguists' concept of Universal Grammar: universal properties of language help learners generalize correctly from their limited experience. It follows that studying universal properties of language can reveal properties of the learner. But the arrow of explanation goes in the opposite direction: if the learner generalizes in particular ways (according to its defining properties), then it can explain some observed universal properties of natural language. In short, a well-articulated theory of language learning offers explanatory adequacy.

This article explores these ideas in the context of phonotactic patterns derived from consonantal harmony patterns. Consonantal harmony patterns are chosen because their typology is well

Material in this article was presented at the 2007 annual meeting of the Linguistic Society of America, at a phonology seminar at the University of Maryland in 2007, and at the Workshop on Language, Cognition, and Computation at the University of Chicago in 2008. I thank the audiences at those presentations for their constructive comments and questions. I also thank the three anonymous reviewers, Daniel Blanchard, Bruce Hayes, Bill Idsardi, Jason Riggle, James Rogers, Edward Stabler, Colin Wilson, Alan Yu, and Kie Zuraw for valuable feedback. This research was supported by a 2008–2009 University of Delaware Research Fund (UDRF) grant.

Linguistic Inquiry, Volume 41, Number 4, Fall 2010
623–661

studied (Hansson 2001, Rose and Walker 2004), so it becomes possible to identify plausible universal properties. Phonotactic patterns are considered instead of alternations for several reasons. First, phonotactic learning has been a recent, active area of research (Hayes and Wilson 2008, Heinz 2009). Second, there is evidence that humans learn phonotactic patterns prior to alternations. Not only are infants sensitive to surface sound patterns (Friederici and Wessels 1993, Jusczyk et al. 1993), but also they use this language-specific phonotactic knowledge to identify word boundaries in speech (Mattys et al. 1999, Mattys and Jusczyk 2001). Thus, it is plausible that acquisition of phonotactic knowledge precedes acquisition of morphophonology. Third, some learning models indicate that phonotactic knowledge is helpful when learning alternations (Albright and Hayes 2002, Hayes 2004, Prince and Tesar 2004). Finally, the phonotactic learning problem is formally simpler than the alternation learning problem because instead of learning a mapping from underlying forms to surface forms, one only has to learn whether surface forms are well formed.

This article shows how specific properties of long-distance phonotactic patterns themselves are sufficient to license correct generalization from only finitely many examples. These properties are the importance of the order of sounds and the relative unimportance of the distance between them. Taken together, these concepts define the notion of precedence (cf. the definition of precedence in Reiss and Mailhot 2007), which is not to be confused with immediate precedence (Raimy 2000).

In a nutshell, the learner acquires a grammar which decides that a word is well formed if and only if its (potentially discontiguous) subsequences of length 2 are well formed. These are exactly the pairs of symbols that stand in the precedence relation. The acquisition of the grammar is straightforward: precedence relations observed in the learner's input are assumed to be well formed and unobserved ones are assumed to be ill formed. The grammars and learner are described in detail in sections 5 and 6, respectively.

The proposed learner, which I call the *precedence learner* because it uses only the notion of precedence to make generalizations, is interesting for many reasons.

1. It is simple, efficient, and provably correct.
2. The patterns learnable by the learner nontrivially approximate the known typology. In particular, the learner explains the presence of two types of attested long-distance phonotactic patterns derived from consonantal harmony patterns (symmetric and asymmetric) and the absence of unattested patterns with blocking effects.
3. The noncounting nature of nonlocal dependencies automatically follows from the way the learner generalizes.
4. The learner lends support to the idea that long-distance phonotactic patterns are phenomenologically distinct from spreading patterns, contra the hypothesis of Strict Locality (Gafos 1999, et seq.).
5. It is the first learner for long-distance dependencies in phonology that does not require an a priori notion of tier or projection (contra the model in Hayes and Wilson 2008).
6. It does not rely on the additional structure provided by Optimality Theory (OT) grammars (Prince and Smolensky 1993, 2004).

7. It does not rely on the additional structure provided by the parameters or cues in the principles-and-parameters (P&P) framework (Chomsky 1981, Dresher and Kaye 1990, Gibson and Wexler 1994).

The precedence learner also raises new questions. As explained in section 6, it is unable to learn phonotactic patterns derived from long-distance dissimilation patterns. Additionally, since this article describes the learner in categorical terms with segmental representations for reasons described in section 3, it remains open how best to extend these results to learners of gradient patterns describable with featural representations.

The proposal here bears some parallels to Search analyses of vowel harmony (Nevins 2005, Reiss and Mailhot 2007, Samuels 2009). There as well as here, the concept of precedence plays a significant role in defining the kinds of patterns phonological grammars can generate and in assessing the minimal computations necessary for describing long-distance dependencies.

This article goes beyond that research in three ways. First, it shows where attested long-distance patterns derived from consonantal harmony fall in the subregular hierarchy, which classifies logically possible patterns according to independently converging measures of computational complexity (McNaughton and Papert 1971, Rogers et al., to appear, Rogers and Pullum, to appear). This exercise makes clear the relevance of order and the irrelevance of distance, which are the defining properties of precedence. Second, this exercise also provides a sound mathematical basis for relating ''precedence as a kind of locality'' (cf. discussion in Reiss and Mailhot 2007) to computational complexity. Third, by demonstrating the contribution an a priori notion of precedence makes to learning, this article factors consonantal harmony patterns in a way that allows researchers to independently investigate the contribution of other factors such as the similarity of sounds.

1.2 The Perspective

This article suggests that phonological learning is modular. One idea is that complex systems that exhibit different kinds of patterns are learned by complex learners: in particular, learners consist of sublearners, and each sublearner learns a submodule of the system. Not only does this idea have proponents who study learning from a biological perspective (Gallistel and King 2010: chap. 13), it is a natural interpretation of significant results from formal learning theories.

Formal learning theories show that learning is impossible unless the range of hypotheses the learner is willing to entertain is a priori restricted (Vapnik 1998, Jain et al. 1999, Niyogi 2006, de la Higuera 2010). The intuition behind these results comes from recognizing that learning is about making distinctions, but not making too many. In a discussion on verb learning, Gleitman (1990:12) sums it up this way: ''The trouble is that an observer who notices *everything* can learn *nothing* for there is no end of categories known and constructable to describe a situation [emphasis in original].'' (See also discussion in Piattelli-Palmarini 1980.)

Under this perspective, Universal Grammar (UG) is the set of available hypotheses, or the properties that these hypotheses share. Shifting focus from the hypothesis space to the learner itself provides explanatory adequacy: properties of the generalization strategy itself ultimately determine the shape of the hypothesis space and hence properties of natural language patterns.

This perspective—that learners are intimately related to their hypothesis spaces—naturally leads to the idea that different kinds of patterns might have different kinds of learners. For example, the set of syntactic hypotheses available to children is not the same as the set of phonological hypotheses available to children. The two domains do not have the same kinds of patterns and so it is reasonable to expect them to be the result of different kinds of learning processes.

The picture that emerges is that the language learner is a composition of mini–learning processes.[1] In the domain of phonotactics, the same thinking carries through. As shown in section 4, the class of long-distance phonotactic patterns measurably differs from the class of local restrictions. It follows that these different classes of patterns could have different learners, of which the complete phonotactic learner is some combination.

Notably, this perspective is largely absent in traditional models in generative grammar like OT or P&P. In those frameworks, the content of the constraints or the parameters determines the typology, but the primary learning proposals are independent of that content (Dresher 1999, Heinz 2009). For example, in OT the content of CON matters not one whit to descriptions of Recursive Constraint Demotion (Tesar and Smolensky 2000:5–6). The learners in these proposals are unable to explain properties of the language patterns themselves because the explanations instead lie in the presence and absence of particular constraints (or parameters).[2]

Here, the claim is that the phonotactic learning module consists of at least one submodule for learning long-distance phonotactics and one for learning local phonotactics. It is plausible that a third module exists for learning surface stress patterns (Heinz 2007, 2009).

1.3 Outline

In section 2, the typological studies of long-distance agreement (LDA) and the relevant issues are reviewed in more detail. Section 3 is intended to provide a framework within which all phonotactic patterns can be described and motivates particular expositional choices. Section 4 explicitly defines long-distance phonotactics (LDP) and classifies them according to the subregular hierarchy (McNaughton and Papert 1971). The results of this query reveal

1. the *importance of order* in LDP and the relative *unimportance of distance*,
2. that a debate about the nature of locality in phonology is also a debate about computational complexity in phonology, and
3. that long-distance dissimilation fundamentally differs from long-distance assimilation with respect to the kinds of phonotactic patterns they derive.

[1] This resembles what formal learning theorists call *parallel learning* (Case and Moelius 2007).

[2] Within OT, there are particular learning proposals that fail to learn the factorial typology and in this way constrain the predicted possible languages (Tesar and Smolensky 2000, Boersma 2003). Hammond (1991) proposes this idea in the context of the P&P framework (and Myers (2002) suggests the similar idea that the subset of the factorial typology reachable by natural processes of sound change are the predicted possible languages). See also Pater 2009 for discussion in the context of Harmonic Grammar. None of these proposals goes as far as the present one in expressing not only the modular-learning perspective but also the natural intimacy that exists between learners and the patterns they learn (the typology).

Section 5 introduces a new class of grammars called *precedence grammars* that are defined in such a way as to transparently capture these insights and that generate a class of patterns called *precedence languages* that includes attested LDP patterns and excludes many unattested ones. Section 6 presents a very simple learner that efficiently learns the class of precedence languages. Section 7 discusses how the precedence grammars can be implemented in gradient frameworks with phonological features, as well as some of the implications the model has for phonological theory—in particular, for nontonal autosegmental tiers and for the role of similarity in LDA.

2 Long-Distance Agreement

2.1 Definition

In their seminal typological studies of consonant harmony, Hansson (2001) and Rose and Walker (2004) define LDA as follows:

> (1) Consonantal long-distance agreement patterns are those where agreement for an articulatory or acoustic property holds between consonants separated by at least one segment. (Rose and Walker 2004:476)

The surveys by Hansson (2001) and Rose and Walker (2004) provide many examples of consonantal LDA, including sibilant harmony like the Navajo example in (2), liquid harmony, dorsal harmony, nasal harmony, and laryngeal harmony, among others. These patterns are not rare. Hansson (2001) documents about 120 languages that require certain consonants to agree in some feature. Hansson (2001) and Rose and Walker (2004) focus exclusively on consonantal harmony, and they leave open the possibility that their analyses may extend to the domain of vowel harmony (see also Hansson 2006).

As an example, consider Navajo, where the anteriority of sibilants within a word is influenced by the anteriority of the rightmost sibilant (Sapir and Hoijer 1967, Fountain 1998). (Data from Sapir and Hoijer 1967:15.)

(2) a. /sì-ʔá/ → sì-ʔá ‘a round object lies’
b. /sì-tí/ → sì-tí ‘he is lying’
c. /sì-ɣìʃ/ → ʃì-ɣìʃ ‘it is bent, curved’
d. /sì-teːʒ/ → ʃi-teːʒ ‘they (dual) are lying’

What is striking about the pattern is how sibilants assimilate despite the arbitrary distances between them. Although Sapir and Hoijer (1967) observe that sibilant harmony in Navajo is less likely to hold as the distance between the sibilants increases, this fact does not change the essential nature of the problem: the anteriority of a sibilant depends in some way on another sibilant that may be arbitrarily distant. A minimally adequate theory that captures this nonlocal dependency must show how one sibilant can affect another across arbitrarily long distances. The observation that the frequency of the effect is reduced as the distance increases does not change this fact.[3]

[3] It is also possible that the distance effect is not a grammatical fact, but due to performance. See related discussion in Hansson 2001:221–223.

Finally, a theory that attempts to mediate the long-distance assimilation through adjacent prosodic domains is inadequate. The sibilants that stand in agreement do not need to be in adjacent syllables, or even in adjacent feet (Hansson 2001, Rose and Walker 2004).

Interestingly, only a few languages with LDA clearly include grammatical restrictions on length (Rose and Walker 2004). These cases suggest that the domain of application of the rule is limited to prosodic domains. For example, in Ndonga, [l] assimilates to [n] if it is separated from a nasal by one vowel but not two vowels (Viljoen 1973). The domain of application of the agreement—syllable, word, or phrase—is an important aspect of the competence of the native speaker. Since this article focuses on the nonlocal character of consonantal harmony and not its domain of application, I set aside the issue of learning the domain of application of the harmony process, and I return to these issues as future work in section 7.

2.2 Spreading, Blocking, and the Typology of Long-Distance Agreement

LDA patterns have been extensively studied by earlier researchers, who have identified many relevant factors (Jensen 1974, Odden 1994, Walker 2000, Hansson 2001, 2007a, Rose and Walker 2004). These factors include the similarity of the sounds that undergo agreement, how the triggering segment is determined, the directionality of the agreement, and the domain of the agreement.

The character of the intervening sounds has also been scrutinized. Hansson (2001:42) adds to the definition in (2) the following:

(3) The intervening segments between the agreeing segments are not audibly affected by the agreeing feature.

The purpose of Hansson's definition is to clearly distinguish LDA patterns known as ''feature spreading''—that is, patterns where arbitrarily long sequences of contiguous segments agree in a feature. The classic example is nasal spreading. For example, in the Johore dialect of Malay, oral vowels and glides may not immediately follow a nasal consonant, nasalized vowel, or nasalized glide (Onn 1980). Consequently, there are words like [peŋãw̃ãsan] 'supervision', but none like *[peŋawasan] or *[peŋãwasan]. Although it is true in [peŋãw̃ãsan] that [ŋ] and the second [ã] agree in the feature [nasal] and are separated by two intervening segments, neither Hansson (2001) nor Rose and Walker (2004) consider this phenomenon to be a case of LDA. This is because the intervening segments participate in the agreement as well.

Some researchers hypothesize that all cases of LDA are feature spreading (with changes to quality of the intervening segments along the relevant phonetic dimension), thus reducing the apparent nonlocal character of the patterns to an unbounded sequence of local transfers. In its strongest form, this means that so-called long-distance patterns are simply an extended kind of coarticulation. This perspective is perhaps most clearly articulated by Gafos's (1999) hypothesis of Strict Locality (see also Ní Chiosáin and Padgett 2001). For Navajo sibilant harmony, this means that all intervening segments between agreeing sibilants realize the feature [anterior]. Thus, the correct representation of 'they (dual) are lying' is not [ʃiteːʒ], but [ʃi̲te̲ːʒ], where the underscore indicates a [−anterior] allophone of some kind, which is audible and perceptible at some level.

A similar, but phonological, analysis posits that the agreeing feature spreads through an unbounded series of local transfers, but it differs from the hypothesis of Strict Locality in that a rule removes the feature from intervening segments with which it is incompatible in a Duke-of-York-style derivation.[4] This rule can be either in the phonetics or late in the phonology, and, unlike under the hypothesis of Strict Locality, the relevant feature is not detectable in the pronounced form of the intervening segments.

In contrast to those approaches, Hansson (2001) and Rose and Walker (2004) argue that LDA patterns are phenomenologically distinct from spreading patterns and consequently require a different kind of analysis. They propose that the agreeing segments are in a particular relationship, which is not shared by the intervening segments. Consequently, as in the phonological analysis above, the intervening segments are not ''audibly affected'' by the agreeing feature.[5] In this respect, their analyses are similar to the Search analysis of Reiss and Mailhot (2007).

A key point in the debate about whether LDA is feature spreading or not is whether long-distance agreement patterns allow blocking. This is because spreading patterns uncontroversially admit blocking effects, regardless of whether all intervening segments in the surface form are ''audibly affected.'' For example, in the Johore dialect of Malay, voiceless obstruents block the nasal spreading (note that nasality stops with [s] in [peŋãw̃ãsan] 'supervision').

When the typology of LDA patterns is considered, both Hansson (2001) and Rose and Walker (2004) observe that

(4) There are no LDA patterns with blocking effects.

They conclude that if feature spreading were the right treatment of LDA, then (4) would be unexpected, and they consequently treat (4) as evidence for their proposals.[6]

However, it remains unresolved what principle, if any, could explain the absence of blocking in the typology of LDA. Rose and Walker's (2004) agreement-by-correspondence (ABC) analysis of LDA in OT has two components. First, there are CC-Correspondence constraints, which place two consonants in correspondence if they are sufficiently similar. The analysis remains agnostic about the exact similarity metric to be used, as this choice is independent of the main thrust of their analysis, which addresses how agreement between segments that are ''similar'' is enforced.[7] The second component of ABC are ID-CC(F) constraints, which enforce agreement of feature F for consonants in correspondence. This analysis is intended to capture both the similarity and blocking effects. However, Hansson (2007a) shows that the ABC approach does predict nonlocal blocking effects of certain types, and he reluctantly suggests that the absence of blocking patterns may be an accidental gap.

[4] This idea survives in OT in turbidity theory (Goldrick 2001). See Finley 2008 for an example of this approach in addressing transparent vowels in vowel harmony.

[5] Hansson and Rose and Walker presumably admit a certain amount of coarticulation on neighboring segments, but not on all intervening ones.

[6] Sanskrit and Kinyarwanda have been proposed as counterexamples. But many researchers diagnose the Sanskrit pattern as feature spreading (Schein and Steriade 1986, Gafos 1999, Ní Chiosáin and Padgett 2001, Hansson 2001, Rose and Walker 2004). Similarly, evidence presented by Mpiranya and Walker (2005), Byrd et al. (2006), and Walker (2007) suggests that Kinyarwanda also displays a feature-spreading pattern.

[7] See also Hansson's (2007b) more recent discussion regarding agreement along secondary consonantal articulations.

There are two additional typological characteristics that Hansson (2001) and Rose and Walker (2004) mention that are relevant to this article.

(5) a. Both symmetric and asymmetric LDA patterns are well attested.
b. The segments participating in the agreement are similar.

The significance of (5b) is discussed in section 7. With respect to (5a), *symmetric LDA* refers to patterns like the one in Navajo where both [− anterior] and [+ anterior] sibilants are ''active'' and force assimilation (here, regressively) to earlier-occurring sibilants. This stands in contrast to asymmetric patterns like the one found in Sarcee. Sarcee also has a sibilant harmony process, but only the [− anterior] sibilants are active, forcing [+ anterior] sibilants to assimilate (again, regressively) (Cook 1978a,b, 1984). Consequently, unlike in Navajo, [+ anterior] sibilants can follow [− anterior] sibilants in Sarcee, though, as in Navajo, the reverse is prohibited. A [+ anterior] sibilant like [z] may follow a [− anterior] sibilant like [ʃ] (as in (6a)), but not vice versa (as in (6c)) (examples (6a–b) from Cook 1978a).

(6) a. /si-tʃiz-aʔ/ → **ʃítʃídz**àʔ 'my duck'
b. /na-s-ɣatʃ/ → nā**ʃ**ɣá**tʃ** 'I killed them again'
c. cf. ****sítʃídz**àʔ

The symmetric/asymmetric types of LDA carry over into LDP as explained in section 3. Importantly, both kinds are well attested. In fact, sibilant harmony patterns, which constitute a major class of consonantal harmony patterns, divide into approximately equal numbers of symmetric and asymmetric forms (Hansson 2001:469–472). Thus, there is no reason to think one kind has a privileged status.

3 Long-Distance Phonotactics

3.1 Motivation

In this article, the studies of Hansson (2001) and Rose and Walker (2004) are recast in terms of phonotactic patterns. In other words, the Navajo pattern is described only in terms of word well-formedness.

(7) Words of Navajo are well formed as long as the sibilants in the word agree in [anterior].

This statement is true, though it says nothing about whether the harmony is regressive or progressive, or root or stem controlled. It says nothing about how the surface form is derived from the underlying form. Instead, the statement in (7) focuses singularly on the fact that a phonotactic pattern can be described as a set of words. All logically possible words that obey the pattern belong to the set, and all logically possible words that do not obey the pattern do not belong to the set.[8] In other words, recasting LDA in terms of long-distance phonotactics (LDP) shifts focus to the weak generative capacity of phonological grammars—that is, to the sets of strings generable

[8] This formulation of pattern (and language) is due to Chomsky (1957).

by the grammars and not the structural descriptions or derivations of those strings (Chomsky 1965:60–62; see also Miller 1999).

This shift in focus is justified for many reasons, some already discussed. The set of strings a grammar weakly generates is a property of the grammar. In the problem of learning phonological grammars, a minimal requirement is that the learned grammar be able to correctly discriminate between logically possible strings that obey the target language pattern and those that do not. While characterizing a grammar's weak generative capacity is a formally simpler problem, it is still nontrivial (Vapnik 1998, Jain et al. 1999, Johnson 2004).

Furthermore, characterizing such strings is plausibly a first step toward learning the grammar itself. Research has shown that infants are sensitive to the surface sound patterns of their language (Friederici and Wessels 1993, Jusczyk et al. 1993). Other research has shown that infants use language-specific phonotactic knowledge to identify word boundaries in speech (Mattys et al. 1999, Mattys and Jusczyk 2001). This research suggests that infants learn phonotactic patterns prior to morphophonological alternations. Additionally, phonological learning models have been proposed in which prior phonotactic knowledge helps the learning of alternations (Albright and Hayes 2002, Hayes 2004, Prince and Tesar 2004, Pater and Tessier 2006).

3.2 Patterns Are Functions

Phonotactic patterns are the rules and constraints that determine the well-formedness of logically possible words (Chomsky and Halle 1965, 1968, Halle 1978, Goldsmith 1994, Heinz 2007, Hayes and Wilson 2008). Therefore, phonotactic patterns can be conceived as functions that map logically possible words to values. Under a categorical phonotactic model, these values could be {0,1} for ''ill formed'' or ''well formed,'' respectively. On the other hand, under a gradient phonotactic model, these values might be the real interval [0,1], where *1* is interpreted as ''most well formed'' and *0* as ''least well formed'' and the intermediate values indicate intermediate levels of well-formedness.

For example, if we model the phonotactics of English as a categorical function, then we might require *English* (slem) = 1, but *English* (srem) = 0 and *English* (pzarʃk) = 0. This function models the observation that English speakers, despite having the same amount of experience with these three hypothetical words (that is to say, no experience), recognize *slem* to be a possible word of English, but neither *srem* nor *pzarʃk*. However, if we model *English* as a gradient function, then the model can make additional distinctions between logically possible words (Albright and Hayes 2003), and accordingly we may desire our function to have the property that *English* (slem) > *English* (srem) > *English* (pzarʃk).

Putting aside the gradient/categorical distinction for a moment, one may consider the function *English* to be some composition of several other phonotactic functions, which pick out the particular phonotactic problems in logically possible strings. For example, *English* may be composed partly from the function **pz*, which maps all logically possible words containing a *pz* sequence to a value less than logically possible words without this sequence. As a categorical function, **pz* maps *slem* and *srem* to 1, and *pzarʃk* and *pzapza* to 0. As a gradient function, **pz* could map

slem and *srem* to 1, *pzarʃk* to 0.1, and *pzapza* to 0.01. The idea that the whole phonotactic function is a composition of other, simpler phonotactic functions is a common idea in generative phonology. For example, it is explicitly expressed in OT (Prince and Smolensky 1993, 2004) as markedness constraints.

Returning to Navajo, whether or not a word is well formed is determined in part by a particular phonotactic constraint that penalizes logically possible words with sibilants with different values of the feature [anterior]. Unlike **pz*, these patterns are usually written with ellipsis points: **s . . . ʃ*. This notation is meant to indicate a function that maps logically possible strings that contain [s] followed somewhere by [ʃ]—no matter what intervenes, or how much—to a lesser value than logically possible strings that do not contain two such segments. For example, as a categorical function, **s . . . ʃ* would map *sotos* and *toʃotoʃ* to 1, *sotoʃ* and *sotoʃotosoʃ* to 0. As a gradient function, **s . . . ʃ* might map *sotos* to 1, *sotoʃ* to 0.1, and *sotoʃotosoʃ* to 0.01.

3.3 Properties of Phonotactic Patterns

When we consider phonotactic patterns as functions, it is natural to ask what kind of function they are. We want to know, out of all the logically possible functions mapping strings to values, what properties make a function a *phonological* one. Additionally, from the perspective of formal learning theories, we are interested in knowing the contribution these properties can make to learning.

One property of phonological functions for which there is a consensus in generative phonology is that they operate over featural representations. In other words, one better captures the constraint of Navajo with the function suggested by the notation *[αanterior] . . . [−αanterior]. The α notation has the advantage of reducing two statements to one. More importantly, the feature [anterior] has the advantage that it allows us to capture in a single statement all the sounds, and only those sounds, that pattern together the same way in Navajo, and to link that behavior to the position of the tongue tip in the oral cavity.

However, it remains unclear what role featural representations play in learning and generalization. On the one hand, they allow for succinct statements as stated above. On the other hand, they provide languages with a phonological alphabet (Calabrese 1988), whose elements each have a unique featural representation. Consequently, learning procedures that generalize over featural representations have a much larger search space than those that make generalizations over segments. Although one recent study suggests that the larger search space can be effectively searched in the case of English onsets (Hayes and Wilson 2008), another suggests that both featural and segmental representations play a role in phonological learning (Albright 2009). For reasons discussed below, I stick to segmental representations, while acknowledging that featural representations and feature-based generalization constitute an important area of continuing and future research (see also section 7.1).

There is another property of phonological functions around which a consensus may be forming. Recently, there has been some discussion in the field of phonology whether phonotactic patterns are best understood as categorical or gradient functions, with a number of researchers

arguing in favor of gradience (Coleman and Pierrehumbert 1997, Coetzee 2008, Hayes and Wilson 2008). From the perspective of learning theory, whether the co-domain is real or categorical matters little to learnability (Vapnik 1998, de la Higuera 2010).[9]

This article has very little, if anything, to say about whether phonotactic patterns are categorical or gradient, nor anything to say about the appropriate use of phonological features in generalization. Instead, it examines other aspects of the nature of phonotactic functions that are orthogonal to the properties mentioned above.

Consider this: regardless of whether we state the constraint as *$s \ldots \int$ or as *[αanterior] . . . [−αanterior], and regardless of whether or not the constraint is categorical or gradient, the fact remains that the quality of some sound in a word depends on the quality of another sound that may be arbitrarily far from it. On the other hand, with constraints like *pz, the quality of some sound in a word depends on the quality of an immediately adjacent sound. Again, this fact is independent of whether *pz is gradient or categorical, or makes use of phonological features in its definition. In other words, the functions *pz and *$s \ldots \int$ are very different in nature regardless of whether we treat them as gradient or categorical, or whether we define them in terms of phonological features. Properties like ''is gradient'' or ''is defined in terms of phonological features'' do nothing to distinguish these functions, and therefore say nothing about some of their essential properties.

What kind of property could distinguish between *pz and *$s \ldots \int$? The Chomsky hierarchy (figure 1), which classifies functions according to the kinds of rewrite grammars that can define them (Chomsky 1956, 1959), provides some possibilities.[10] The finite languages are properly included by the regular languages, which are properly included by the context-free languages, which are properly included by the context-sensitive languages.[11] Uncontroversially, whether a function is gradient, is categorical, or makes use of phonological features has zero impact on its place in the hierarchy.

To illustrate, students of formal language theory learn that the pattern $a^n b^n$ represents the set of strings in (8), where ϵ indicates the unique string of length 0 (the ''empty'' string).

(8) {ϵ, ab, aabb, aaabbb, . . . }

This set is essentially the same as a categorical function that maps the strings in (8) to 1 and logically possible strings not in (8) to 0.[12] Interestingly, there is no regular (i.e., right-branching)

[9] Space does not permit discussion, but Horning's (1969) results and work that supersedes Horning's (Osherson, Weinstein, and Stob 1986, Angluin 1988) are about whether the sequence of data presented to the learner has specific properties. In fact, Gold (1967) presents a similar result. See the discussion in Heinz and Riggle, to appear.

[10] It is typically said that the Chomsky hierarchy classifies languages, or patterns, as opposed to functions, but *patterns are functions.* See footnote 12.

[11] These regions can be defined in terms other than rewrite grammars. One common way to define them is in automata-theoretic terms; see Hopcroft, Motwani, and Ullman 2001 or Kracht 2003.

[12] Generally, any set can be conceived as a categorical function and vice versa. The characteristic function, also called the indicator function, of the set suffices. Let S be any set. The indicator function $f_S(x)$ is 1 iff $x \in S$ and 0 otherwise.

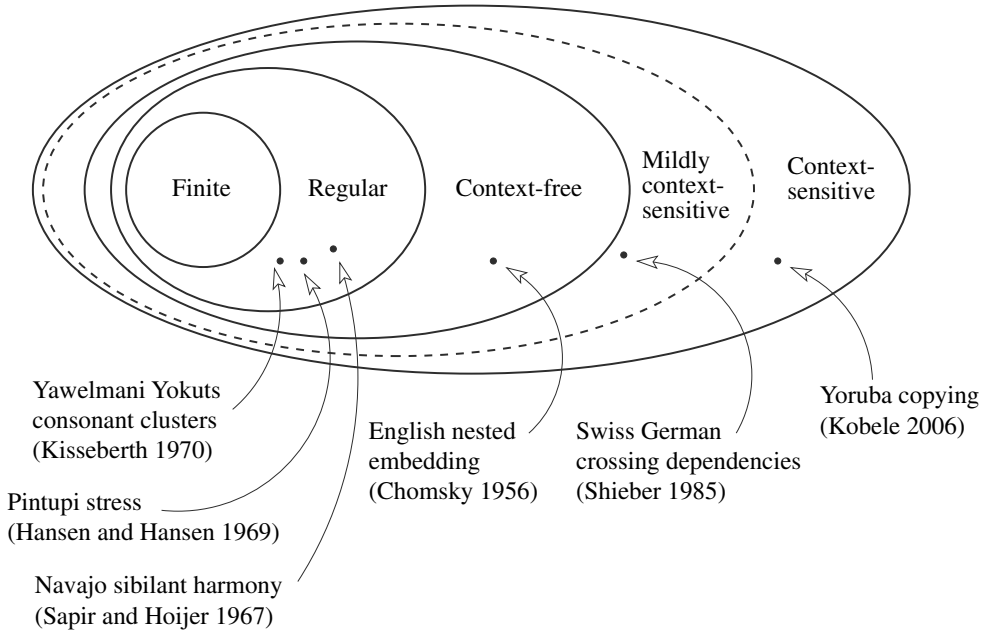

Figure 1
The Chomsky hierarchy

rewrite grammar that generates the set in (8), but there is a context-free rewrite grammar that generates it, shown in table 1. Although the textbook example is categorical, it is easy to make any rewrite grammar gradient by assigning probabilities to the rules of the grammar (for details see, for example, Roark and Sproat 2007 or Kornai 2007). No amount of gradience can make the function recognizing $a^n b^n$ regular. Likewise, replacing *a* and *b* with features indicating some natural class does not change this fact. The property ''is context-free'' is completely orthogonal.

The constraints **pz* and **s . . . ʃ*, unlike the pattern $a^n b^n$, are *regular*. The property ''is regular,'' like ''is context-free,'' is a property independent of gradience, categoricality, or the use of phonological features. Regular patterns are those that can be described with right-branching

Table 1
Categorical and gradient context-free grammars for $a^n b^n$

Categorical	Gradient	
S → ϵ	S → ϵ	(0.5)
S → aSb	S → aSb	(0.5)

Table 2
Rewrite-grammar representations of *pz and *$s \ldots \int$

*pz		*$s \ldots \int$	
S → vS	S → ϵ	S → cS	B → cB
S → cS	A → ϵ	S → vS	B → vB
S → zS		S → sA	B → ʃB
S → pA		S → ʃB	
A → cS		A → cA	S → ϵ
A → vS		A → vA	A → ϵ
A → pA		A → sA	B → ϵ
v = any vowel		v = any vowel	
c = any consonant except {p, z}		c = any consonant except {s, ʃ}	

rewrite grammars, or equivalently, with finite-state automata.[13] Right-branching rewrite grammars are those in which the right-hand side of every production rule either has the form xC (x an alphabetic symbol and C a category) or is the empty string ϵ.

Table 2 shows regular rewrite grammars for *pz and *$s \ldots \int$. It is easy to verify, for example, that the grammar for *pz accepts all and only those strings with no pz sequence in them. It cannot accept any string with a pz sequence because in order to generate the symbol p, the grammar must then use a production rule whose left-hand side consists of A. But no such rule generates the symbol z, so the grammar rejects all strings with a pz sequence. The rewrite grammar for *$s \ldots \int$ is more complex (it has more rules and one more category), but it is still regular.[14] It is worthwhile to convince oneself that this grammar also generates all and only those strings that do not contain both s and $\int$ (no matter how far apart they are in a word). To conclude, properties like ''is regular'' can distinguish between functions like *pz and *$s \ldots \int$ on the one hand and $a^n b^n$ on the other (since no right-branching rewrite grammar can describe $a^n b^n$), but properties like ''is gradient'' or ''is described with phonological features'' cannot.

Much attention has been drawn to the fact that syntactic patterns are at least context free, even context sensitive (Chomsky 1956, 1959, Shieber 1985, Kobele 2006) (see figure 1). Although it has been known for over thirty-five years, less attention has been paid to the fact that phonological patterns are overwhelmingly *regular* (Johnson 1972, Koskenniemi 1983, Ellison 1994, Kaplan and Kay 1994, Eisner 1997, Albro 1998, 2005, Frank and Satta 1998, Karttunen 1998, Gerdemann and van Noord 2000, Riggle 2004, Heinz 2007, 2009).[15] In other words, although sentence well-formedness patterns seem to necessitate (mildly) context-sensitive computations over words, word well-formedness seems only to require regular computations over individual sounds.[16] For example, to my knowledge, no one has described any phonotactic pattern that remotely resembles $a^n b^n$.

[13] See Kracht 2003:chap. 2 for additional characterizations of regular sets.
[14] The rewrite grammar in table 2 describes both *$s \ldots \int$ and *$\int \ldots s$.
[15] For dissenting views, see Barton, Berwick, and Ristad 1987 and Reiss 2009.
[16] The notable exception to this is reduplication, which is arguably a morphological process (Inkelas and Zoll 2005). However, there are finite-state approaches to reduplication (Roark and Sproat 2007).

On the other hand, cooccurrence restrictions like *pz are commonplace, as are long-distance patterns like consonantal harmony. These patterns are all regular (as shown in section 4).

Although the property ''is regular'' appears to be sufficient to distinguish phonotactic patterns from syntactic patterns, in many ways it is insufficient as a characterization of phonotactic patterns in general. In the first place, being regular does not distinguish *pz from *$s \ldots ʃ$, as they are both regular. Also, even though all phonotactic patterns are regular, there are many regular patterns that make for unattested and unnatural phonotactic patterns. For example, the pattern that requires a string to have an even number of consonants and an even number of vowels, regardless of their order, is a regular pattern. Words like *baba*, *bbaa*, *aabb*, *baab* are all well formed according to this pattern, unlike words like *baa*, *aba*, *ababa*, *bba*. So the property ''is regular'' is not a sufficient condition to make a pattern a phonotactic one. Finally, there is another sense in which the property ''is regular'' is insufficient. The property ''is regular'' characterizes a class of languages too large and unstructured for a learner to exactly acquire a regular language from positive evidence only (Gold 1967). The learning issues are discussed in section 6.

To summarize this section: While we are ultimately interested in all properties of phonological functions, featural representations and properties like gradience are irrelevant to the place of functions in the Chomsky hierarchy, which classifies functions according to other kinds of properties. Far from being tangential to the concerns of phonologists, these other kinds of properties are as central to the pursuits of phonological theory as these others, if not more so. Knowing whether a function is gradient or categorical tells us little about its nature. It tells us nothing about whether a function picks out a cooccurrence restriction like *pz, a long-distance dependency like *$s \ldots ʃ$, or a more complex function like *$a^n b^n$. The same is true for featural representations. One question this article answers is what property determines whether a pattern is like *pz or like *$s \ldots ʃ$. In both cases, the notion of order matters, but in the latter case, the notion of distance does not.

Because the gradient/categorical distinction and the issue of phonological features are orthogonal to this fundamental aspect of the long-distance patterns investigated here, the main exposition of this article treats these patterns as categorical. Similarly, the description of the learner in section 6 makes no mention of phonological features. These expositional choices are intended to focus attention on the properties this article claims make phonotactic patterns long-distance and learnable from surface forms alone. Section 7.1 shows how the basic ideas are compatible with gradient functions that make reference to features. Some may argue that stripping away important properties of phonotactic patterns such as gradience and features is misleading or unhelpful, but in fact the opposite is true. By factoring the problem down to its essentials, one sees it for what it is.

4 Long-Distance Phonotactics and the Subregular Hierarchy

This section makes clear the importance of order and the relative unimportance of distance in attested LDP patterns. Together these properties define *precedence*, which is at the heart of the learning proposal. These properties are revealed through discussion of specific examples of patterns relevant to the discussion of LDP. These patterns are cooccurrence restrictions like *pz, spreading patterns, symmetric LDP, asymmetric LDP, LDP with blocking, and LDP derived from

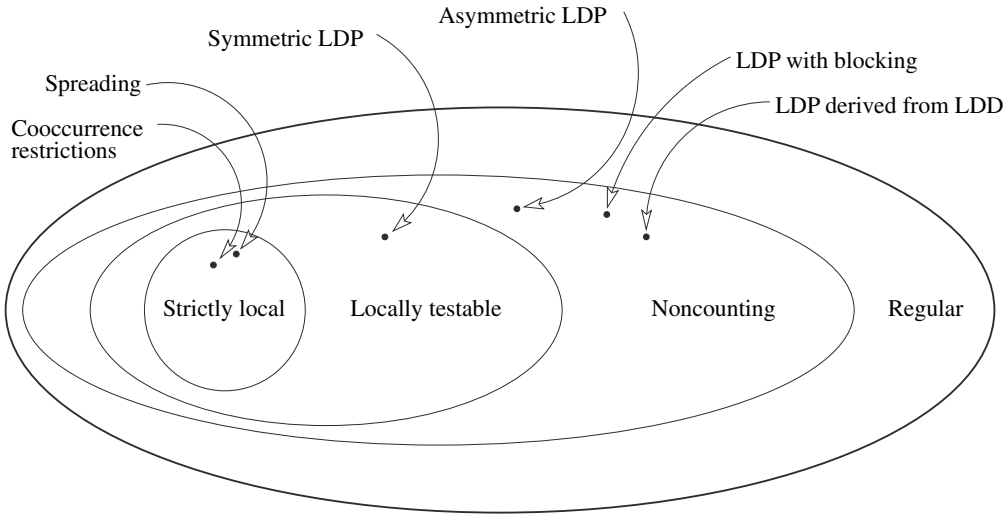

Figure 2
The subregular hierarchy

long-distance dissimilation. These six patterns are introduced now so that it is later clear which patterns the learner can acquire and which it cannot.

These six patterns are presented in the context of the subregular hierarchy, which—like the Chomsky hierarchy—organizes functions according to independently converging measures of complexity (McNaughton and Papert 1971). Rogers and Pullum (to appear) provide an accessible introduction to the subregular hierarchy and make the case that it provides fertile ground for investigating the auditory pattern recognition abilities of different species.

A diagram of the hierarchy with the locations of the relevant patterns is given in figure 2. The three main classes are the strictly *k*-local languages, the locally testable languages, and the noncounting languages.

4.1 Cooccurrence Restrictions

Cooccurrence restrictions are patterns that assign lower well-formedness values to logically possible words that contain some illegal contiguous sequence than to those logically possible words that contain no instances of the illegal contiguous sequence. For example, the categorical function **pz* assigns a value of ''well formed'' to all the words in (9) since they have no *pz* sequence, and ''ill formed'' to all other words.

(9) {kip, slem, srem, t, p, ppppp, . . . }

Cooccurrence restrictions belong to the lowest class in the subregular hierarchy, which are called the strictly *k*-local patterns.[17] Basically, a strictly *k*-local pattern is one in which word well-

[17] This class is also called locally *k*-testable in the strict sense (McNaughton and Papert 1971).

formedness can be determined by considering whether contiguous substrings of length k are well formed. Such substrings are called k-factors in formal language theory.[18]

For example, the 2-factors of *pzarʃk* are {*pz*, *za*, *ar*, *rʃ*, *ʃk*}. If these are all allowed, then *pzarʃk* is well formed. On the other hand, if *pz* is not on the list of allowable 2-factors, then *pzarʃk* is ill formed. Here is another example: if the allowable 2-factors are {*ab*, *bb*, *ba*}, then the pattern consists of all words ''constructable'' from those 2-factors, like those shown in (10).

(10) {*ab*, *bb*, *ba*, *abb*, *abba*, *bba*, *bbb*, . . . }

Every string occurs in (10) except those that contain the 2-factor *aa* (or any other 2-factor not in the allowable list). Since **pz* can be described by a list that includes all 2-factors except for *pz*, it is a strictly 2-local pattern.[19]

In terms of phonotactics, this means that a phonotactic pattern is strictly k-local if the well-formedness of a word can be determined solely by checking whether the substrings of length k are allowable. The constraint *CCC (Kisseberth 1970) is a strictly 3-local pattern. Many phonotactic patterns are strictly k-local, though there is some question as to the upper bound of k (see Heinz 2007, Hayes and Wilson 2008 for related discussion).

If there is a k such that a pattern is strictly k-local, it is said to be strictly local. It can be shown that if a pattern is strictly k-local, it is also strictly $(k + 1)$-local (McNaughton and Papert 1971). Thus, there is an infinite hierarchy of strictly local patterns, beginning with strictly 1-local patterns and going all the way up.

Readers familiar with the field of natural language processing may recognize that strictly k-local grammars are similar to n-gram grammars (the n-gram is a contiguous subsequence of length n; i.e., a k-factor where $k = n$). They are in fact the same, the difference being that n-gram grammars typically represent probability distributions over all possible words (i.e., are gradient functions), whereas strictly k-local ones are categorical. Often n-gram grammars are given as charts like the one in table 3. The probability of a string is determined by multiplying probabilities associated with each of the n-grams in a word. We can also represent a strictly k-local grammar in this way, where the weights are now just {0,1} since the grammar is categorical.

If k is known in advance, strictly k-local patterns are learnable by a very simple kind of learner (Garcia, Vidal, and Oncina 1990, Heinz 2007). The procedure is very similar to the training of n-gram models employed in natural language processing tasks (Manning and Schütze 1999, Jurafsky and Martin 2000). Essentially, these algorithms keep track of the observed k-factors (or

[18] Technically, the strictly k-local languages are allowed to specify particular $k - 1$ factors at the left and right edges of the word (essentially the same as considering word boundaries as part of the word). This detail is ignored here for easier exposition.

[19] Technically, for some alphabet Σ, we let a grammar G be some subset of Σ^2. The language of the grammar $L(G)$ is defined as the set $\{w$: the k-factors of w are a subset of $G\}$. It is logically equivalent to define the grammar as a list of *prohibited* k-factors, in which case the language of the grammar consists of all words whose k-factors are not found in this grammar. The choice made here is an expositional one, which hopefully makes it easier to understand the proposed learner. See section 6.2 for further discussion.

Table 3
n-gram and strictly *k*-local grammars

n-gram grammar				
	p	z	. . .	a
p	0.25	0.01		0.25
z	0.25	0.25		0.25
⋮			⋱	
a	0.25	0.25		0.25

Strictly *k*-local grammar				
	p	z	. . .	a
p	1	0		1
z	1	1		1
⋮			⋱	
a	1	1		1

n-grams) in the linguistic environment (e.g., corpus). These learners are guaranteed to succeed provided the linguistic experience is sufficient (Garcia, Vidal, and Oncina 1990, Heinz 2007). For example, if the first word a categorical learner encounters is the word *abba*, the learner can infer that the grammar includes the 2-factors {*ab*, *bb*, *ba*} and hence generalizes to the language given in (10). The precedence learner discussed in section 6 is very similar.

4.2 Symmetric Long-Distance Phonotactics

Navajo sibilant harmony is an example of symmetric LDP. Recall that in Navajo, a word is not well formed if it contains sibilants that disagree in the feature [anterior]. The words in (11a–b) obey this pattern; the words in (11c–d) do not.

(11) a. ʃiteːʒ ‘they (dual) are lying’ c. *ʃiteːz
b. dasdoːlis ‘he (4th) has his foot raised’ d. *dasdoːliʃ

Descriptively, this kind of pattern can be expressed with the following two statements (I forego the [α] notation to contrast Navajo with Sarcee below):

(12) 1. [−anterior] sibilants never precede [+anterior] sibilants.
2. [+anterior] sibilants never precede [−anterior] sibilants.

Thus, hypothetical words such as *sotos* and *ʃotoʃ* are possible words of Navajo. On the other hand, words like *ʃotos* and *sotoʃ* are not possible words of Navajo. This pattern is symmetric because both statements in (12) are required to describe the pattern in full.

Long-distance phonotactic patterns like symmetric LDP are not strictly local for any *k*. This is because the material that separates the agreeing segments can be arbitrarily long, and therefore longer than *k*. Symmetric LDP patterns belong to the next rung in the subregular hierarchy, the locally testable class.

A locally *k*-testable pattern is one that can be obtained under the Boolean closure of the strictly *k*-local class.[20] It turns out that a pattern is locally *k*-testable if and only if it is possible

[20] Thus, any strictly *k*-local pattern is also locally *k*-testable, but not vice versa.

to decide whether a word obeys the pattern simply by considering whether the *set of* k-*factors* making up the word is allowable. Consequently, any locally 2-testable pattern either includes both *fifizt* and *fififizt* or excludes both (since they have the same set of *k*-factors: {*fi*, *if*, *iz*, *zt*}). Unlike strictly *k*-local languages, however, a locally *k*-testable pattern may include a word like *rakt* but exclude a word like *rak* since the two words have *different* sets of *k*-factors. (A strictly *k*-local language that includes *rakt* must allow *k*-factors {*ra*, *ak*, *kt*} and therefore must also include *rak*.)

Symmetric LDP patterns like Navajo are locally 1-testable. This is because in order to decide whether a hypothetical word is a well-formed word of Navajo, one need only consider the set of 1-factors, that is, the *set of segments* that make up the word. If this set of segments contains two sibilants with different values of [anterior], then the word is not well formed in Navajo (e.g., *sotoʃ* yields {*s*, *t*, *o*, *ʃ*}). If all the sibilants in the set agree in [anterior], then it passes the sibilant harmony test for Navajo well-formedness (e.g., *sotos* yields {*s*, *t*, *o*}). In a sense, we consider every segment in the word *as being adjacent to every other segment*—the distance between the segments does not matter at all.

Any locally *k*-testable language is also locally $(k + 1)$-testable but not vice versa (McNaughton and Papert 1971). It follows that there is an infinite hierarchy of locally *k*-testable languages. If there is a *k* such that a pattern is locally *k*-testable, it is said to be locally testable.

Also, for fixed *k*, locally *k*-testable patterns are learnable by a very simple learner, provided again the linguistic environment is sufficient. The learner simply keeps track of the *sets of* k-*factors* found in well-formed strings in the learning environment. In other words, the row headings in a chart like those in table 3 would be subsets of the alphabet. Because of the (much) larger hypothesis space,[21] such learners are not efficient. In fact, there are no efficient learners for this hypothesis space (José Sempere, pers. comm.).

4.3 Asymmetric Long-Distance Phonotactics

Another Athabaskan language provides a useful example of asymmetric LDP. Sarcee, like Navajo, has a sibilant harmony pattern. However, unlike the Navajo pattern, in Sarcee only the [−anterior] sibilants are ''active'' (Cook 1978a,b, 1984). In terms of LDA, this means that only the [−anterior] sibilants regressively change the [+anterior] sibilants (in the underlying form) to [+anterior] sibilants (in the surface form). But the reverse never happens, yielding surface forms like the ones in (6). Descriptively, we can describe the phonotactic pattern as follows:

(13) 1. [+anterior] sibilants never precede [−anterior] sibilants.

Stated this way, the asymmetric LDP pattern is clearly just ''one-half'' of the symmetric LDP pattern of Navajo. Table 4 summarizes Navajo and Sarcee LDP.

Asymmetric LDP patterns are not locally testable, for any *k*. Intuitively, this is because the kind of test above for Navajo will not work for Sarcee. The hypothetical words *ʃotos* and *sotoʃ*

[21] There are $2^{|A|}$ subsets of a set *A*.

Table 4
Symmetric and asymmetric LDP

Hypothetical word	Navajo (symmetric LDP)	Sarcee (asymmetric LDP)
sotos	√	√
ʃotoʃ	√	√
ʃotos	x	√
sotoʃ	x	x

have the same set of 1-factors {*s*, *t*, *o*, *ʃ*}, but only *ʃotos* is well formed. To determine whether a word is well formed in Sarcee, it is important to know something about the *order* of the sibilants.

The noncounting class is formed by closing the locally testable patterns under concatenation and Boolean operations. McNaughton and Papert (1971) show that a pattern is noncounting if there is a number n such that for all strings u, v, w, if uv^nw occurs in L, then $uv^{n+1}w$ occurs in L as well.[22] It turns out that all locally testable languages are noncounting but not vice versa (McNaughton and Papert 1971).

Asymmetric LDP patterns are noncounting patterns. An example of a logically possible, unattested, regular phonotactic pattern that is not noncounting is one in which well-formed words must have an even number of vowels.

4.4 Spreading

Under the hypothesis of Strict Locality (Gafos 1999), all cases of LDA are reduced to spreading operations. Spreading patterns, like the nasal spreading pattern of the Johore dialect of Malay (mentioned in section 2) are strictly 2-local. It is possible to determine whether a word like [peŋãw̃ãsan] 'supervision' obeys the nasal spreading rule by simply checking that each 2-factor in the word is allowed. Since each member of the set {*pe*, *eŋ*, *ŋã*, *ãw̃*, *w̃ã*, *ãs*, *sa*, *an*} is permissible in the language, this word is well formed. On the other hand, a hypothetical word like [peŋawasan] is not well formed in the language since the 2-factor *ŋa* is not permitted. Note that, in this simple model, the fact that voiceless obstruents block the spreading of nasality is captured, in part, by allowing the 2-factors {*ãs*, *sa*}.

Under the hypothesis of Strict Locality, it follows that there are no cases of ''real'' LDP, since the sibilant harmony phonotactic can be described with a strictly 2-local language. It is an interesting coincidence that Gafos's term *Strict Locality* matches the formal language theory term so well. Gafos's hypothesis of Strict Locality solves the problem of long-distance phonotactics by eliminating them entirely. They do not exist, as all such patterns are actually strictly 2-local. Furthermore, under this hypothesis, if one looks carefully enough, one ought to be able to detect the [anterior] feature on the intervening segments (Gafos 1999). (See also Gordon 1999, Ní Chiosáin and Padgett 2001, Gafos and Benus 2003, and Gick et al. 2006.)

[22] This class goes by many names; others that may be familiar include *locally testable with order* and *star-free*.

This hypothesis is contrary to the one offered by Hansson (2001) and Rose and Walker (2004), who maintain that the agreement is genuinely long-distance, without the support of intervening segments overtly carrying the relevant feature. From the vantage point offered by the subregular hierarchy, one sees this debate is partly about how complex phonological patterns are, with Gafos on the side of ''simpler'' and Hansson and Rose and Walker on the side of ''more complex.''

4.5 Long-Distance Phonotactics with Blocking

On the basis of their typological studies, Hansson (2001) and Rose and Walker (2004) independently arrive at the same conclusion: LDA with blocking is unattested. This feature of the typology carries over to LDP. In other words, LDP patterns derived from LDA patterns, like the LDA patterns themselves, do not exhibit blocking patterns. To see this, consider what an LDP pattern with blocking would look like.

Imagine a language such that voiceless sibilants must agree in the feature [anterior] unless there is a voiced sibilant intervening between two disagreeing voiceless sibilants. For example, like Navajo, this language permits words like *sotos* and *ʃotoʃ* and prohibits words like *ʃotos* and *sotoʃ*. In other words, the constraints *$s \ldots ʃ$ and *$ʃ \ldots s$ appear to be in effect. However, unlike Navajo, this language also permits words like *soztoʃ* and *ʃotozos* because the voiced sibilant blocks the agreement. Informally, we might write the constraints as *$s \ldots ʃ$ *(z)* and *$ʃ \ldots s$ *(z)*, where the symbol in parentheses indicates a blocking element. Table 5 provides some additional examples of well-formed and ill-formed words according to this pattern.

Where does this pattern fall within the subregular hierarchy? Like asymmetric LDP, it is noncounting. Similarly to Sarcee, a word like *sozoʃ* obeys the LDA-with-blocking pattern above but *ʃosoz* does not. Yet they have the same set of 1-factors, {*ʃ, s, o, z*}. Generally, for any *k*, we can find two words—one that obeys the pattern and one that does not—that have the same set of *k*-factors. Consequently, this pattern cannot be locally testable and is in fact noncounting.

4.6 Long-Distance Phonotactics Derived from Long-Distance Dissimilation

Both Hansson (2001) and Rose and Walker (2004) also deliberately exclude long-distance dissimilation (LDD) from their studies (see Hansson 2001:5–6). This leaves open the question whether LDP derived from LDA differs in any significant way from LDP derived from LDD.

Table 5
Well- and ill-formed words of a hypothetical LDP-with-blocking pattern

Well-formed words	Ill-formed words
ʃotoʃ	*ʃotos
sotos	*sotoʃ
ʃozos	*ʃosozos
sosozoʃ	*soʃosozos

When LDD is translated to a phonotactic pattern, it looks like LDP with blocking.[23] Consider long-distance liquid dissimilation found in Latin (Jensen 1974, Odden 1994).[24] In Latin, [l] becomes [r] if another [l] occurs earlier in the word.[25]

(14) a. /nav-alis/ nav-alis 'naval'
b. /episcop-alis/ episcop-alis 'episcopal'
c. /infiti-alis/ infiti-alis 'negative'
d. /sol-alis/ sol-aris 'solar'
e. /lun-alis/ lun-aris 'lunar'
f. /milit-alis/ milit-aris 'military'

However, the rule does not apply if an [r] intervenes.

(15) a. /flor-alis/ flor-alis 'floral' *flor-aris
b. /sepulkr-alis/ sepulkr-alis 'funereal' *sepulkr-aris
c. /litor-alis/ litor-alis 'of the shore' *litor-aris

It is possible to interpret the behavior of the suffix as a response to the phonotactic constraint (16).

(16) In well-formed words, [l]s never precede [l]s unless [r] intervenes.

Stated this way, LDD looks similar to hypothetical LDP with blocking, discussed in section 4.5. In the informal notation used here, we could abbreviate (16) as **l . . . l (r)*. Like the hypothetical case above, this pattern is noncounting.

There is another interpretation of the data. Perhaps the constraints in Latin include both **r . . . r* and **l . . . l*, and a form like *flor-alis* 'floral' is the optimal form even though it violates **l . . . l*. Such an analysis requires one to explain why no modification results in an improvement. A partial answer might be found in that **flor-aris* violates **r . . . r*.

It is striking that LDP derived from LDA differs significantly from LDP derived from LDD. In the former case, blocking effects are absent and the generalizations are exceptionless. In the latter case, either there are regular exceptions to the generalization, or the generalization must be stated as a kind of blocking pattern. This article offers no account of long-distance phonotactic patterns with blocking. For more on this issue, see section 6.3.

4.7 The Typology of Long-Distance Phonotactics

Classifying the phonotactic patterns above in terms of the subregular hierarchy is revealing. First, it makes concrete the aspect of LDP that *ignores distance* (recall the Navajo test above). Also, it reveals a significant difference between symmetric and asymmetric LDP, the relevance of *order*. Third, it shows the typological characteristics of LDA described in (4) and (5), transfer to the

[23] Thanks to Alan Yu for bringing cases like this to my attention.
[24] See Suzuki 1998 for many other cases of LDD.
[25] A reviewer points out some exceptions, such as *fili-alis* 'filial' and *glute-alis* 'gluteal'.

domain of phonotactics. Fourth, it shows that the debate about locality in phonology is partly a debate about the computational complexity of phonological patterns. Fifth, it reveals that phonotactic patterns obtained from long-distance *dissimilation* patterns may be a kind of LDP with blocking. Sixth, the subregular hierarchy is not fine-grained enough to distinguish between robustly attested asymmetric LDP, unattested LDP with blocking, and rarely attested LDD, all of which are noncounting patterns.

5 Precedence Grammars and Languages

This section introduces a class of formal languages I call *precedence languages*, which properly include symmetric and asymmetric LDP to the exclusion of LDP-with-blocking patterns. Thus, this class closely approximates the attested patterns. Precedence languages are those languages that can be generated by a certain kind of grammar, called a *precedence grammar*. Precedence grammars underlie models of reading comprehension (Whitney and Berndt 1999, Whitney 2001, Grainger and Whitney 2004, Schoonbaert and Grainger 2004) and some models of text classification (Shawe-Taylor and Cristianini 2005:chap. 11), though to my knowledge, apart from Heinz 2007 and Rogers et al., to appear, this is the first time they have been made explicit. Rogers et al. refer to this class as the strictly 2-piecewise languages, because they are precisely analogous to the strictly 2-local languages. Rogers et al. show that the strictly piecewise patterns form the basis of another subregular hierarchy with multiple independent, converging characterizations.

The driving concept behind precedence grammars is that order matters, but not distance. To see this clearly, reconsider the sibilant harmony pattern of Navajo.

(17) In well-formed words, sibilants must agree in the feature [anterior].

As mentioned in section 4.2, (17) can be translated into the following two statements:

(18) 1. [−anterior] sibilants never precede [+anterior] sibilants.
2. [+anterior] sibilants never precede [−anterior] sibilants.

These two statements can be made into a longer, even more explicit list.

(19) [s] can be preceded by [s].
[s] can be preceded by [t].
. . .
[t] can be preceded by [s].
. . .
[ʃ] can be preceded by [ʃ].
[ʃ] can be preceded by [t].
. . .

The crucial features of the list in (19) are that (a) it is finite and (b) it contains no statements where a [αanterior] sibilant precedes a [−αanterior] sibilant.

The list in (19) is a *precedence grammar*. A precedence grammar is a list of the allowable *precedence relations* in a language. The precedence relations in a word *w* are the pairs of segments

Table 6

A precedence grammar for a fragment of Navajo

$$G = \begin{Bmatrix} ss & & st & so \\ & ʃʃ & ʃt & ʃo \\ ts & tʃ & tt & to \\ os & oʃ & ot & oo \end{Bmatrix}$$

xy that make the statement ''*x* precedes *y* in *w*'' true. The language of a precedence grammar *G* consists of all words *w* such that the precedence relations in *w* are also in *G*.[26]

For example, consider the precedence grammar in table 6. The language of this grammar includes a word like *sotos* because the precedence relations present in the word *sotos*—that is, each pair of {*so*, *st*, *ss*, *ot*, *os*, *to*, *ts*}—is in the grammar. Similarly, it can be seen that the language of *G* includes *ʃotoʃ*. However, a word like *ʃotos* does not belong to the language of *G* since the precedence relation *ʃs* is not in *G*. Essentially, the precedence grammar *G* encodes a fragment of the Navajo grammar in (18), where the only sibilants are *s* and *ʃ*.

The exercise above is familiar. It is the same kind of process used to determine whether a pattern is strictly 2-local. There, one checks the 2-factors of a word against some allowable list. The allowable list is essentially the grammar of the strictly 2-local language. Here, instead of checking the 2-factors of the word against some list, one checks the precedence relations in the word.

One might wonder if the grammar could be described more compactly if instead of statements like ''x can be preceded by y'' the grammar included statements like ''x cannot be preceded by y,'' and if languages were defined to be all and only those words whose precedence relations were not in the grammar. For the attested cases under consideration, such grammars would be smaller to state. However, with respect to the kinds of languages that can be defined one way or the other, the two ways are equivalent. Since expressivity is the focus here, and since the first way makes the learning procedure in section 6 easier to describe, the discussion continues with the above definitions.

The example in (19) shows that symmetric LDP can be described with precedence grammars. It is easy to see that precedence grammars can also describe asymmetric patterns. Adding the precedence relation *ʃs* to the grammar in (19), for example, defines a grammar that is essentially a fragment of Sarcee. This is because this language accepts, like Sarcee, words in which [s] follows [ʃ], but it continues to prohibit [ʃ] following [s].[27]

It is also easy to see that no LDP-with-blocking pattern is a precedence language. This is simply because precedence grammars admit no concept of blocking. If a language like the hypothetical language in section 4.5 contains the word *sozoʃ*, then the precedence grammar must

[26] Formally, for some alphabet Σ, we let a grammar *G* be some subset of Σ^2 (the allowable precedence relations). Then the language of the grammar $L(G)$ is defined as the set $\{w : \text{the precedence relations of } w \text{ are a subset of } G\}$. Note the similarity to the formal definition of strictly 2-local languages. See footnote 19.

[27] The asymmetric property of precedence relations is discussed in Raimy 2000 and Reiss and Mailhot 2007.

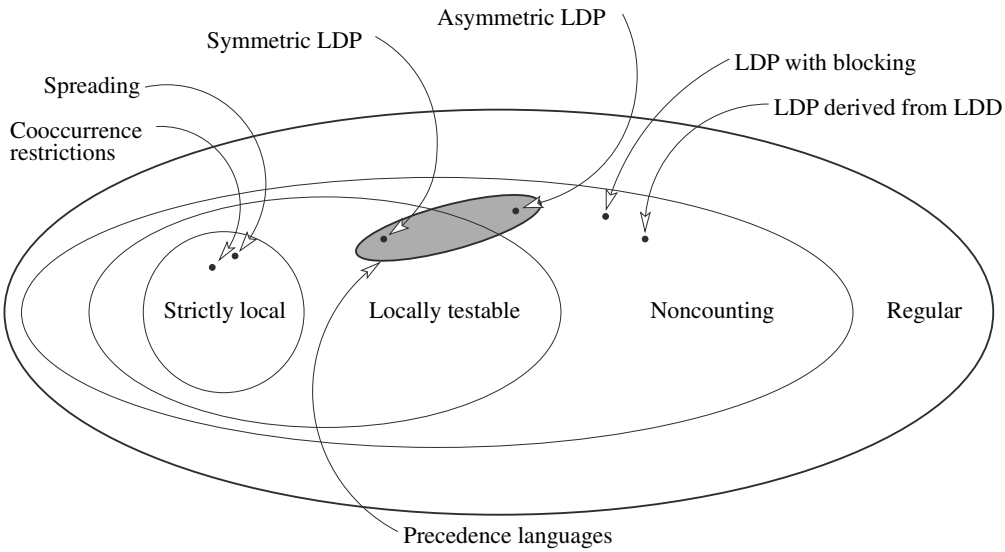

Figure 3
Precedence languages in the subregular hierarchy

include the precedence relations *sf*, *so*, *of*. Hence, this language must also include the word *sof* and cannot be the hypothetical pattern discussed in section 4.5.[28]

It follows that long-distance phonotactic patterns derivable from long-distance dissimilation patterns are also not describable with precedence grammars. This is true, and it remains an open question how precedence grammars can be generalized to include such patterns.

As shown in figure 3, precedence languages carve out a small region that crosscuts the subregular hierarchy. For details about their properties and the piecewise subregular hierarchy, see Rogers et al., to appear.

6 Learning Long-Distance Phonotactics

This section introduces the learning mechanism that acquires LDP patterns, of the kind found in Navajo and Sarcee, from finitely many examples. This learner is evaluated in the Gold (1967) learning framework known as *exact identification in the limit from positive data*, and it is shown that this learner provably learns the precedence languages in this framework. The Gold framework is chosen because it focuses attention on generalization. As discussed later, the way the learner

[28] Extending precedence relations to triples *abc*, which means ''[a] precedes [b], which precedes [c],'' does not make it possible to account for blocking. Consider the word *sozof*. While it is true that *szf* is a discontiguous subsequence of length 3 in this word, so is *sof*. Consequently, the language of this grammar would also include a word like *sof*, which disobeys this LDP-with-blocking pattern.

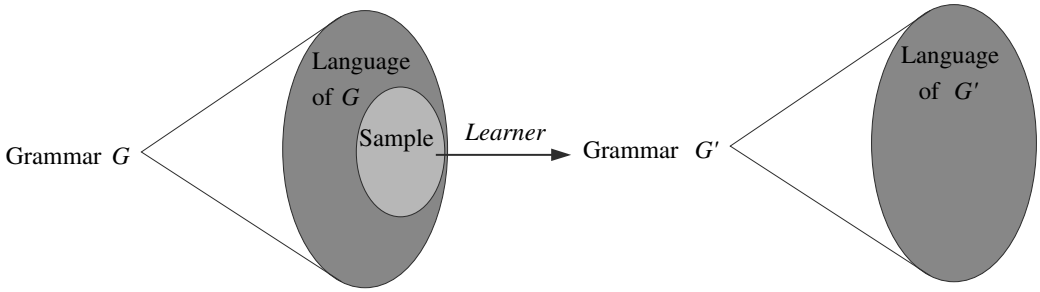

Figure 4
The learning framework

generalizes transfers successfully to other learning frameworks, which make different assumptions—for example, with respect to the learning environment or the success criteria.

6.1 Learning Patterns

It is useful to make clear the learning framework, schematized in figure 4. The idea is that the language the learner is trying to learn is generated from some grammar G. The learner, however, does not hear every element of the language (as it is infinite in size); rather, it hears only some small finite sample. The learner is a function that maps finite samples to grammars.

The central question of interest is, What is the learner such that language of $G' =$ language of G? A learner successfully learns a language L if, upon being presented with ever larger finite samples from L, the grammars returned by the learning function converge to one that generates exactly L. See Nowak, Komarova, and Niyogi 2002 for an accessible introduction to this framework and Osherson, Weinstein, and Stob 1986, Jain et al. 1999, Niyogi 2006, and de la Higuera 2010, for more technical introductions.

A central result is that there is no learner for any class of languages that properly includes all finite languages (Gold 1967). Hence, there is no learner that can learn the regular, context-free, or context-sensitive language classes. However, classes that crosscut these regions that exclude some finite languages may be identifiable in the limit from positive data provided they have the right properties (Angluin 1980), and the grammatical inference community has identified many such classes (de la Higuera 2005, 2010).

One critique of this framework is that the input to the learner is too generous. The learner is guaranteed to see any finite collection of words from L that it may need.[29] Also, the input to the learner is nonnoisy. For these reasons, the framework does not faithfully resemble the real-life experience of language-learning humans. This critique is a misunderstanding of the purpose

[29] More precisely, the input to the learner is an infinite text, and every word in L occurs at least once in this text. The learner must converge to the target language after seeing some finite portion of this text.

of this framework, which is to illuminate how learners can generalize beyond their finite experience in the first place (see Nowak, Komarova, and Niyogi 2002 and Johnson 2004 for more discussion).

6.2 Learning Precedence Grammars

This section presents a very simple kind of learner, which can learn precedence languages. This learner generalizes by making distinctions with respect to precedence but not distance. It follows that if the properties of the learner are taken to be basic, then the essential properties of the class follow from the basic design of the learner.

The initial state of the learner's precedence grammar is empty. All the learner does is record the precedence relations in observed words. For example, table 7 shows how the grammar grows when the learner observes the sequence *tosos*, *ʃotoʃ*, *stot*. New precedence relations added to the grammar are given in boldface type. On the basis of these few forms, the learner already generalizes tremendously. It accepts words like *ʃoʃ*, *ʃtot*, and *sototos* but not words like *ʃtos* or *sosoʃ*. Provided that the words heard by the learner were generated by a speaker of the grammar given in table 6, which only generates words obeying the sibilant harmony constraint, then there are no words observable by the learner that could add any precedence relations to the grammar after the last time step in table 7.

In this way, a learner that records precedence relations observed in the sample identifies the language of the grammar *G* in the limit because all the precedence relations in the target grammar are present in only finitely many forms. In fact, it can be shown that, at each time step, the precedence learner hypothesizes the smallest precedence language consistent with the sample observed so far (Heinz 2007). Thus, convergence is guaranteed for any LDP pattern.

The precedence learner closely resembles the learner for strictly 2-local languages, discussed in section 4.1. In fact, the only difference is that strictly 2-local learners pay attention to immediate

Table 7
Precedence learning: Navajo sibilant harmony

Time	Word	Precedence relations	Grammar
0			Ø
1	tosos	to, ts, os, oo, so, ss	**ss** **so** **ts** **to** **os** **oo**
2	ʃotoʃ	ʃo, ʃt, ʃʃ, ot,oo, oʃ, to, tʃ	ss **ʃʃ** **ʃt** so **ʃo** ts **tʃ** to os **oʃ** **ot** oo
3	stot	st, so, to, tt, ot	ss **st** so ʃʃ ʃt ʃo ts tʃ **tt** to os oʃ ot oo

precedence (i.e., 2-factors or contiguity relations), whereas precedence learners pay attention to general precedence.[30] In Heinz 2010, I show that these two learners are specific instances of a general kind of learning strategy called *string extension learning*, and Kasprzik and Kötzing (2010) provide an elegant generalization of string extension learning in terms of lattices.

Finally, let us return to the alternative way to define precedence grammars: as a list of prohibitive precedence relations. The languages these grammars generate include all and only those words whose precedence relations are not prohibited by the grammar. Under this construal, the learning procedure is as follows. The initial state of the grammar is the complete list of precedence relations. As words are observed, the precedence relations that are present in those words are *removed* from the grammar. Although this procedure is different, the learner, as a function from samples to languages, is the same. For example, with the first word *tosos* above, the learner would remove the precedence relations {*to*, *ts*, *os*, *oo*, *ss*} from the grammar. Consequently, at this point the language the learner recognizes, like the learner illustrated in table 7, includes words like *sosososos* and *tooooos*. So although there are procedural differences in the descriptions of these two learners, they are functionally identical.

6.3 Explaining the Absence of Long-Distance Phonotactics with Blocking

Precedence-based learners cannot learn LDP with blocking. This straightforwardly follows from the observation that precedence learners only acquire precedence grammars, which describe precedence languages, which do not include LDP-with-blocking patterns. However, it is instructive to run through an example.

Recall the hypothetical LDP-with-blocking pattern from section 4.5. (See table 5 for examples of well- and ill-formed words from this hypothetical language.)

When exposed to a word from this language such as *ʃozos*, the precedence learner overgeneralizes irreparably. This is because the language of the learner's grammar now includes words like *ʃos*, since *ʃs*, *ʃo*, *os* are precedence relations in *ʃozos* and are thus entered as part of the grammar. There is simply no way for the precedence learner to acquire LDP-with-blocking patterns. As explained in footnote 28, redefining precedence grammars as sets containing triples (x, y, z) does solve this problem.

A similar argument shows that LDP derived from LDD patterns are not acquirable by this learner, either. This is true regardless of whether one describes the phonotactics as a pattern with blocking or without it. For example, if the phonotactic in Latin is **l . . . l (r)*, then as described above the learner will fail. Even if the target constraints are **r . . . r* and **l . . . l*, the learner fails because of opaque forms like *litor-alis* 'of the shore' (since the learner will infer that the precedence relation *ll* is part of the grammar).

The latter scenario suggests one research possibility: if the learner is modified to handle noise, then perhaps it could acquire **l . . . l* in the face of counterexamples like *litor-alis* 'of the

[30] This similarity has led Grainger and Whitney (2004) to call precedence relations ''open'' bigrams. The term *precedence relation* appears to more transparently reveal the essential nature of the model. The term *2-length subsequence* is perhaps most accurate (Simon 1975, Rogers et al., to appear).

shore'. If successful, this modification would allow a unified analysis of LDP derived from LDA and LDP derived from LDD. The difference between the two types comes down to whether their optimal forms violate a phonotactic constraint or not. In LDP derivable from LDA, optimal forms do not violate the long-distance phonotactic (e.g., Navajo *ʃitɛɮʒ* 'they (dual) are lying' does not violate either **ʃ . . . s* or **s . . . ʃ*); but in LDP derivable from LDD, optimal forms do (e.g., Latin *litor-alis* 'of the shore' violates **l . . . l*).

Pursuing this line of analysis is outside the scope of this article. However, if nothing else, the proposal here reveals an interesting difference between LDP patterns derivable from LDA and LDP patterns derivable from LDD: namely, the former can be identified in the limit from positive data by precedence learners, and the latter cannot.

Nonetheless, when one's attention returns to LDP derived from LDA, the conclusion is clear. If humans generalize in the way suggested by the precedence learner, it explains why such patterns fail to exhibit blocking effects.

6.4 Efficiency of the Precedence Learner

The learning procedure outlined above, which I call precedence learning, is tractable. This is because the number of precedence relations in a word is given by a quadratic function in the length of the word. Furthermore, this function is bounded from above by a constant: the square of the number of alphabetic symbols. Since computing the precedence relations in a string is quadratic in the length of the string, learning time is quadratic in the size of the sample.

Additionally, it is also possible to characterize the finite samples that are sufficient for successful learning. A sample is sufficient provided, for every precedence relation in the target grammar, there is some word in the sample with this precedence relation. In the example in table 7, {*tosos*, *ʃotoʃ*, *stot*} is a sufficient sample since every precedence relation in the grammar in table 6 is in the sample.

Also, the size of a sufficient sample for a precedence language grows polynomially with the size of the grammar (see de la Higuera 1997 for some discussion). Note that the sample size is relatively small. In fact, it is on the order of $|\Sigma|^2$, where $|\Sigma|$ is the number of symbols in the alphabet. This is no different from the size of the sample needed to learn strictly 2-local languages.

6.5 Modular Language Learning

The precedence learner cannot learn cooccurrence restrictions like **pz*. But it does not have to. Such constraints can already be learned by a strictly 2-local learner. The proposal here is that language learning is modular. The complete phonotactic learner consists of (at least) two learning modules: one for learning cooccurrence restrictions, and one for learning long-distance constraints.

For concreteness, imagine the final phonotactic grammar consisting of a strictly 2-local grammar and a precedence grammar. How does a complex grammar like this determine word well-formedness? Ideally, the overall well-formedness of a word would be compositionally related to the well-formedness score returned by the two subparts of the total grammar. In a categorical model, perhaps the simplest method is to require a word to be well formed if and only if it is well formed according to both components of the grammar. This is just the intersection of the

acquired strictly 2-local language and the precedence language. This is not the only way well-formedness scores from the component grammars could be composed to determine an overall score, but I leave such issues as a matter of future research.

6.6 Local Summary

To summarize, learning long-distance phonotactic patterns like the symmetric and asymmetric ones in Navajo and Sarcee is easy. As is discussed in further detail below, it is easy without a notion of a sibilant tier, or even a notion of similarity. It is easy because precedence-based learners do not consider every logically possible nonlocal environment. Although they make distinctions based on order, precedence-based learners do not distinguish on the basis of distance at all. Consequently, precedence-based learners cannot learn logically possible nonlocal patterns like these:

(20) a. If the third segment after a sibilant is a sibilant, they must agree in [anterior].
b. If the second, third, or fifth segment after a sibilant is a sibilant, they must agree in [anterior].
c. . . .

The fact that these kinds of patterns are all unattested follows directly from how the precedence learner generalizes. In other words, the generalization strategy proposed here explains the absence of logically possible counting patterns like those in (20). In addition to failing to learn counting patterns, precedence-based learners cannot learn LDP-with-blocking patterns. It is striking that making (and not making) these distinctions is not only sufficiently restrictive for learning to occur, but also allows both symmetric and asymmetric LDP to be learned.

It is reasonable to conclude that if humans generalize in the way suggested by the precedence learner, it explains why

1. long-distance consonantal harmony patterns do not count intervening segments,
2. there are symmetric LDP patterns,
3. there are asymmetric LDP patterns, and
4. LDP derivable from LDA exhibit no blocking patterns.

Finally, as a corollary to the above, it follows that LDP patterns are distinct from spreading patterns, which commonly exhibit blocking effects (which follows from their status as a strictly 2-local type of pattern).

7 Discussion

7.1 Extending Precedence-Based Learning

Throughout this article, I used a categorical model with segmental representations instead of featural ones. I did so in order to facilitate exposition of the central point: that the notion of precedence, and the generalizations obtainable from this notion, solves some of the puzzles and illuminates some of the issues surrounding LDP, and consequently LDA.

Now, however, the time has come to address these issues, and to show that these assumptions can be relaxed with little difficulty and are therefore not as problematic as they might initially seem. The purpose here is not to examine any particular proposal in great detail—that is beyond the scope of this article, whose main points have already been established. Instead, it only remains to show that there is every reason to believe that it is possible to extend the precedence-based learner so that it accommodates similarity and learns gradient grammars.

7.1.1 Learning Gradient Phonotactics There are many ways to make gradient precedence-based learners. Perhaps the most straightforward way is to estimate the probabilities using the finite-state structure of precedence languages (Heinz and Rogers 2010).

Another possibility is to plug these patterns into Hayes and Wilson's (2008) maximum entropy model. In the same way that this model currently discovers *n*-gram-style constraints and assigns them weights, it can do the same for precedence-style constraints.

7.1.2 Phonological Features Precedence grammars stated over featural representations dramatically increase the learner's search space because there are many more natural classes than segments (see Hayes and Wilson 2008 for related discussion), potentially making it difficult to make efficient computations. It remains an open question for the field of computational linguistics and grammatical inference how the structure of the feature system may help a search through this space (see Albright 2009 and open question #4 in de la Higuera 2006), though Heinz and Koirala (2010) offer one solution.

Hayes and Wilson's (2008) approach suggests another way phonological features could be integrated into precedence learning: their model could just as easily search for prohibitive precedence relations stated over phonological feature bundles in addition to searching for *n*-gram constraints.

7.1.3 Learning the Domain of Agreement It was assumed throughout the earlier discussion that the domain of application of the agreement is known and that the learner receives domain-sized units. In most cases of LDA, the relevant domain appears to be the word. Only in a few languages is the domain of application of the agreement smaller than the word (Rose and Walker 2004). Is it possible to learn both the domain of the phonotactic and the phonotactic simultaneously? Recent research suggests that phonotactic patterns over words can be obtained when the learning data consist of utterances without word boundaries (Blanchard and Heinz 2008, Blanchard, Heinz, and Golinkoff 2010). This suggests that similar techniques could simultaneously learn the phonotactics over domains smaller than the word (i.e., syllable or foot) while at the same time determining what those domains are.

7.2 Similarity

Precedence grammars can describe patterns that require sound *x* not to follow sound *y*, where *x* and *y* are dissimilar. Thus, to my knowledge, precedence grammars are the first theory in phonology capable of describing LDP patterns without using similarity to license the long-distance dependency. It follows that in its purest form, precedence-based learning predicts that such patterns

ought to be learnable, a conclusion that runs counter to the typological observation that LDA holds between similar sounds (5b).[31]

The typology as far as I know is sound, but it is interesting to ask whether any grammarians would even notice, for example, a static pattern in a language like ''No [ʒ]s follow [b]s.'' Certainly, whether or not adults or children can learn patterns like this is testable in artificial language experiments (e.g., Chambers, Onishi, and Fisher 2002, Onishi, Chambers, and Fisher 2003, Wilson 2006, Cristià and Seidl 2008).

But even if it is shown (as I expect) that people learn long-distance dependencies between similar-sounding segments more easily than between dissimilar ones, thus supporting the typological character of LDP, this is no reason to reject precedence-based learning. Precedence-based learners are independent of any particular theory of similarity in phonology, but they are not incompatible with them.

The simplest proposal is that similarity is an independent filter (possibly with an explanation in speech planning (Hansson 2001, Rose and Walker 2004)), which further restricts the hypothesis space. Thus, the only precedence relations that can be omitted from a grammar are those that meet the similarity criterion (whatever it is). Thus, grammars are not allowed to omit a precedence relation like *zb*, but could omit *sʃ* since [z] and [b] are not sufficiently similar, but [s] and [ʃ] are.

This idea can be implemented either categorically, as outlined above, or gradiently. One possibility is to set the priors in a Bayesian learning model to favor similar-sounding pairs of segments. In this manner, not only does the model operate over the precedence-based hypothesis space, it can be substantially biased toward patterns that are phonetically motivated (Wilson 2006).

In short, there is every reason to be optimistic that a theory that accommodates all of the typological features in (4) and (5) can be developed. Finally, in my opinion, it is an advantage of the approach taken here that similarity is not required for learning long-distance dependencies. A theory that requires fewer assumptions to achieve its goals is to be preferred over one that requires more assumptions. In fact, we can say that we have successfully factored the problem: by keeping track of order, and not distance, long-distance dependencies of a particular sort can be learned. Similarity is now a distinct factor to be studied separately.

7.3 Tiers

The proposal made here, like the Search proposals (Nevins 2005, Reiss and Mailhot 2007, Samuels 2009), does not require an independent theory of phonological tiers. After its introduction as an insightful means for analyzing tonal patterns (Goldsmith 1976), the notion of a phonological tier has played a major role in understanding nonlocal segmental interactions (McCarthy 1979, 1986, Poser 1982, Archangeli and Pulleyblank 1989, 1994). Although tiers for every feature have been

[31] The approach here is probably most similar to Pulleyblank's (2002) approach, which does allow constraints of the form $*x \ldots y$ operating on a particular tier.

proposed (e.g., what Hayes (1990) calls the ''bottlebrush'' theory), most proposals aim to constrain what counts as a tier in some principled way (e.g., Clements 1985, Sagey 1986, Mester 1988).

Recent approaches to learning long-distance phonotactic patterns employ phonological tiers (Ellison 1991, Hayes and Wilson 2008, Goldsmith and Xanthos 2009, Goldsmith and Riggle, to appear). The idea here is straightforward: the Navajo word *ʃiteːʒ* 'they (dual) are lying' has the representation in (21): a melodic tier and, for example, a strident tier.

(21)

strident	ʃ				ʒ			
	↑				↑			
melodic	**ʃ**	i	t	eː	**ʒ**			'they (dual) are lying'

Stridents on the melodic tier project to the strident tier but nonstridents do not. On the strident tier, the segments [ʃ] and [ʒ] are essentially adjacent, as no stridents intervene between them. Another way of putting it is, *nonstridents aside*, the distribution of stridents can be described with a strictly 2-local pattern. I will refer to approaches like this one as *tier-based bigram learners*.

Tier-based bigram learners require an independent theory of tiers. This theory could provide the tiers antecedently, as is explicitly assumed in the learners of Ellison (1994) and Hayes and Wilson (2008). Or this theory could aim to learn them, perhaps along the lines proposed by Goldsmith and Xanthos (2009).

Whatever the theory of tiers is, it ought to be clear that tier-based bigram learners depend on its existence. On the other hand, precedence-based learning requires no independent theory of tiers. Again, all things being equal, a theory with fewer assumptions is to be preferred over one with more assumptions.

As is often the case, however, all things are not equal. One characteristic of tier-based bigram learners is that they can learn LDP-with-blocking patterns. This is because strictly 2-local patterns uncontroversially admit blocking effects. To see this, consider a word from the hypothetical LDP-with-blocking pattern discussed earlier.

(22)

strident	s				z		ʃ	
	↑				↑		↑	
melodic	**s**	o	t	o	**z**	o	**ʃ**	(hypothetical)

If the tier-based bigram learner obtains a strident tier, then it should have no difficulty learning this LDP-with-blocking pattern. At the level of representation of the strident tier, the distribution of the segments is describable exactly with a strictly 2-local pattern—that is, bigrams. The relevant fragment of this grammar is {*ss*, *ʃʃ*, *sz*, *zs*, *ʃz*, *zʃ*}.

It follows that tier-based bigram learners can learn LDP patterns derivable from LDD. For example, in Latin, assuming that a liquid tier has been either discovered or given, the liquid dissimilation pattern can be described by the strictly 2-local grammar {*rl*, *lr*, *rr*} over this tier. This is in contrast to what we have seen with precedence-based learning, which cannot learn any LDP with blocking.

If a bona fide case of LDP (or LDA) with blocking is discovered, this will be a point in favor of tier-based bigram learning approaches. Without this result, it is more difficult to decide

which situation is more desirable. Tier-based approaches to learning can learn all attested patterns, as well as an unattested class of patterns. On the other hand, precedence-based learners learn LDP derivable from LDA but not LDP derivable from LDD. As mentioned in section 4.6, perhaps the constraints in LDD cases are best described with constraints like $*l \ldots l$, but the optimal surface forms (by some calculation) violate such constraints (e.g., Latin *litor-alis* 'of the shore'). The issue is not likely to be settled soon, but I think the fact that precedence-based learning is not contingent upon an independent theory of tiers is a distinct advantage, even if it is unable at present to account for the learnability of LDP derived from LDD.

8 Conclusion

A learner that generalizes with respect to the order of sounds, but ignores the distance between sounds, can learn attested LDP patterns efficiently and easily, from surface forms alone. The learner does not require independent theories of similarity or of phonological tiers (contra the model in Hayes and Wilson 2008).

Importantly, the precedence learner fails to learn LDP patterns with blocking. This provides a reason why LDP cases derivable from LDA cases never exhibit blocking: such patterns, even if they exist in the learning environment, cannot be discovered by the learner. It follows that if people generalize in the manner suggested, it explains not only this aspect of the typology, but also the nature of the attested types of LDP. As noted, such an explanation is typically not available in OT or P&P frameworks since the character of the typology is divorced from current learning proposals (Dresher 1999, Heinz 2009).

This result also lends support to the hypothesis that LDP patterns are phenomenologically distinct from feature-spreading patterns, contra Gafos 1999. This is because feature-spreading patterns uncontroversially admit blocking effects, whose absence from the typology of LDA now has a principled basis.

Precedence learners also cannot learn cooccurrence restrictions. This is indicative of the modular organization of the whole phonotactic-learning mechanism. At a minimum, the complete phonotactic learner requires one module for learning cooccurrence restrictions (strictly local patterns) and another module for learning long-distance patterns (precedence patterns, strictly piecewise patterns). This result follows in part from a research methodology that aims to factor the learning problem by understanding the contribution individual factors can make to learning. Other recent work that factors the learning problem includes the work reported in Heinz 2009 and Tesar, to appear.

In addition to the research directions suggested in section 7, there are several open questions of interest. First, it remains to be seen how the ideas presented here can be adapted to learn patterns of alternations as opposed to patterns over sets of strings. Second, it would be interesting to see where phonotactic patterns derivable from vowel harmony patterns fall in the subregular hierarchy, and whether any fall into the class of precedence languages. Third, it would be interesting to know whether language acquisition experiments confirm whether humans actually generalize in the way suggested by the precedence learner.

References

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26: 9–41.

Albright, Adam, and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, 58–69. Somerset, NJ: Association for Computational Linguistics.

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90:119–161.

Albro, Dan. 1998. Evaluation, implementation, and extension of primitive Optimality Theory. Master's thesis, UCLA, Los Angeles, CA.

Albro, Dan. 2005. A large-scale, LPM-OT analysis of Malagasy. Doctoral dissertation, UCLA, Los Angeles, CA.

Angluin, Dana. 1980. Inductive inference of formal languages from positive data. *Information Control* 45: 117–135.

Angluin, Dana. 1988. Identifying languages from stochastic examples. Technical Report 614, Yale University, New Haven, CT.

Archangeli, Diana, and Douglas Pulleyblank. 1989. Yoruba vowel harmony. *Linguistic Inquiry* 20:173–217.

Archangeli, Diana, and Douglas Pulleyblank. 1994. *Grounded Phonology.* Cambridge, MA: MIT Press.

Barton, G. Edward, Robert Berwick, and Eric Sven Ristad. 1987. *Computational complexity and natural language*. Cambridge, MA: MIT Press.

Blanchard, Daniel, and Jeffrey Heinz. 2008. Improving word segmentation by simultaneously learning phonotactics. In *Proceedings of the Conference on Natural Language Learning*, ed. by Alexander Clark and Kristina Toutanova, 65–72. East Stroudsburg, PA: Association for Computational Linguistics. Available at http://www.aclweb.org/anthology/W/W08-2109.pdf.

Blanchard, Daniel, Jeffrey Heinz, and Roberta Golinkoff. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. In *Computational models of child language learning*, ed. by Brian MacWhinney, special issue, *Journal of Child Language* 37:487–511.

Boersma, Paul. 2003. Review of Tesar & Smolensky (2000): *Learnability in Optimality Theory. Phonology* 20:436–446.

Byrd, Dani, Fidèle Mpiranya, Sungbok Lee, Celeste DeFreitas, and Rachel Walker. 2006. The articulation of consonants in Kinyarwanda's sibilant harmony. Handout of poster presented at the meeting of the Acoustical Society of America, Honolulu, HI.

Calabrese, Andrea. 1988. Towards a theory of phonological alphabets. Doctoral dissertation, MIT, Cambridge, MA.

Case, John, and Sam Moelius. 2007. Parallelism increases iterative learning power. In *Algorithmic Learning Theory: 18th International Conference, ALT 2007*, ed. by Marcus Hutter, Rocco A. Servidio, and Eiji Takimoto, 49–63. Berlin: Springer-Verlag.

Chambers, Kyle E., Kristine H. Onishi, and Cynthia Fisher. 2002. Learning phonotactic constraints from brief auditory experience. *Cognition* 83:B13–B23.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* IT2(3):113–124.

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

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

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, Noam. 1981. *Lectures on government and binding.* Dordrecht: Foris.

Chomsky, Noam, and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1:97–138.

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

Clements, G. N. 1985. The geometry of phonological features. *Phonology Yearbook* 2:225–252.
Coetzee, Andries W. 2008. Grammaticality and ungrammaticality in phonology. *Language* 84:218–257.
Coleman, John S., and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, 49–56. Somerset, NJ: Association for Computational Linguistics.
Cook, Eung-Do. 1978a. Palatalizations and related rules in Sarcee. In *Linguistic studies of native Canada*, ed. by Eung-Do Cook and Jonathan Kaye, 19–35. Vancouver: University of British Columbia Press.
Cook, Eung-Do. 1978b. The synchronic and diachronic status of Sarcee g^y. *International Journal of American Linguistics* 4:192–196.
Cook, Eung-Do. 1984. *A Sarcee grammar.* Vancouver: University of British Columbia Press.
Cristià, Alejandrina, and Amanda Seidl. 2008. Phonological features in infants' phonotactic learning: Evidence from artificial grammar learning. *Language, Learning, and Development* 4:203–227.
Dresher, Elan. 1999. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry* 30:27–67.
Dresher, Elan, and Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* 34:137–195.
Eisner, Jason. 1997. Efficient generation in primitive Optimality Theory. In *Proceedings of the 35th Annual ACL and 8th EACL*, 313–320. Madrid. Available at http://www.aclweb.org/anthology/P/P97/P97-1040.pdf.
Ellison, T. Mark. 1991. The iterative learning of phonological constraints. Ms., University of Western Australia, Perth.
Ellison, T. Mark. 1994. Phonological derivation in Optimality Theory. In *COLING 94*, vol. 2, 1007–1013. Kyoto, Japan. Available at http://www.aclweb.org/anthology/C/C94/C94-2163.pdf.
Finley, Sara. 2008. The formal and cognitive restrictions on vowel harmony. Doctoral dissertation, Johns Hopkins University, Baltimore, MD.
Fountain, Amy. 1998. An Optimality Theoretic account of Navajo prefixal syllables. Doctoral dissertation, University of Arizona, Tucson.
Frank, Robert, and Giorgio Satta. 1998. Optimality Theory and the generative complexity of constraint violability. *Computational Linguistics* 24:307–315.
Friederici, Angela, and Jeanine Wessels. 1993. Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics* 54:287–295.
Gafos, Adamantios. 1999. *The articulatory basis of locality in phonology.* New York: Garland.
Gafos, Adamantios, and Stefan Benus. 2003. On neutral vowels in Hungarian. In *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 3–9 August 2003*, ed. by Maria-Josep Solé, Daniel Recasens, and Joaquín Romero, 77–80. Available at https://files.nyu.edu/ag63/public/papers/gafos-icphs.pdf.
Gallistel, C. R., and Adam Philip King. 2010. *Memory and the computational brain.* Singapore: Wiley-Blackwell.
Garcia, Pedro, Enrique Vidal, and José Oncina. 1990. Learning locally testable languages in the strict sense. In *Algorithmic Learning Theory, First International Workshop, ALT '90*, ed. by Setsuo Arikawa, S. Goto, Setsuo Ohsuga, and Takashi Yokomori, 325–338. Springer/Ohmsha.
Gerdemann, Dale, and Gertjan van Noord. 2000. Approximation and exactness in finite state Optimality Theory. In *Proceedings of the Fifth Meeting of the ACL Special Interest Group in Computational Phonology*, 34–45. Somerset, NJ: Association for Computational Linguistics.
Gibson, Edward, and Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25:407–454.
Gick, Bryan, Douglas Pulleyblank, Fiona Campbell, and Ngessimo Mutaka. 2006. Kinande vowel harmony. *Phonology* 23:1–20.
Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition* 1:3–55.
Gold, E. M. 1967. Language identification in the limit. *Information and Control* 10:447–474.

Goldrick, Matthew. 2001. Turbid output representations and the unity of opacity. In *NELS 30: Proceedings of the 30th Annual Meeting of the North East Linguistic Society*, ed. by Masako Hirotani, Andries Coetzee, Nancy Hall, and Ji-yung Kim, 231–245. Amherst: University of Massachusetts, Graduate Linguistic Student Association.

Goldsmith, John. 1976. Autosegmental phonology. Doctoral dissertation, MIT, Cambridge, MA.

Goldsmith, John. 1994. A dynamic computational theory of accent systems. In *Perspectives in phonology*, ed. by Jennifer Cole and Charles Kisseberth, 1–28. Stanford, CA: CSLI Publications.

Goldsmith, John, and Jason Riggle. To appear. Information theoretic approaches to phonological structure: The case of Finnish vowel harmony. *Natural Language and Linguistic Theory.*

Goldsmith, John, and Aris Xanthos. 2009. Learning phonological categories. *Language* 85:4–38.

Gordon, Matthew. 1999. The ''neutral'' vowels of Finnish: How neutral are they? *Linguistica Uralica* 35: 17–21.

Grainger, Jonathan, and Carol Whitney. 2004. Does the huamn mnid raed wrods as a wlohe? *Trends in Cognitive Science* 8:58–59.

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

Hammond, Michael. 1991. Parameters of metrical theory and learnability. In *Logical issues in language acquisition*, ed. by Iggy Roca, 47–62. Dordrecht: Foris.

Hansen, Kenneth, and Lesley E. Hansen. 1969. Pintupi phonology. *Oceanic Linguistics* 8:153–170.

Hansson, Gunnar. 2001. Theoretical and typological issues in consonant harmony. Doctoral dissertation, University of California, Berkeley.

Hansson, Gunnar. 2006. Locality and similarity in phonological agreement. Talk handout from ''Phonology Fest,'' Indiana University, Bloomington.

Hansson, Gunnar. 2007a. Blocking effects in agreement by correspondence. *Linguistic Inquiry* 38:395–409.

Hansson, Gunnar. 2007b. On the evolution of consonant harmony: The case of secondary articulation agreement. *Phonology* 24:77–120.

Hayes, Bruce. 1990. Diphthongisation and co-indexing. *Phonology* 7:31–71.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: The early stages. In *Fixing priorities: Constraints in phonological acquisition*, ed. by René Kager, Joe Pater, and Wim Zonneveld, 158–203. Cambridge: Cambridge University Press.

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

Heinz, Jeffrey. 2007. The inductive learning of phonotactic patterns. Doctoral dissertation, UCLA, Los Angeles, CA.

Heinz, Jeffrey. 2009. On the role of locality in learning stress patterns. *Phonology* 26:303–351.

Heinz, Jeffrey. 2010. String extension learning. In *Proceedings of the 48th Annual Meeting of the ACL.* Uppsala, Sweden, 11–16 July 2010. Association for Computational Linguistics.

Heinz, Jeffrey, and Cesar Koirala. 2010. Maximum likelihood estimation of feature-based distributions. In *Proceedings of the Eleventh Meeting of the ACL Special Interest Group in Computational Morphology and Phonology (SIGMORPHON)*, 28–37.

Heinz, Jeffrey, and Jason Riggle. To appear. Learnability. In *The Blackwell companion to phonology*, ed. by Marc van Oostendorp, Keren Rice, and Colin Ewen. Blackwell.

Heinz, Jeffrey, and James Rogers. 2010. Estimating strictly piecewise distributions. In *Proceedings of the 48th Annual Meeting of the ACL.* Uppsala, Sweden, 11–16 July 2010. Association for Computational Linguistics.

de la Higuera, Colin. 1997. Characteristic sets for polynomial grammatical inference. *Machine Learning* 27:125–138.

de la Higuera, Colin. 2005. A bibliographical study of grammatical inference. *Pattern Recognition* 38: 1332–1348.

de la Higuera, Colin. 2006. Ten open problems in grammatical inference. In *Grammatical inference: Algorithms and applications. 8th International Colloquium, ICGI 2006, Tokyo, Japan*, ed. by Yasubumi Sakakibara, Satoshi Kobayashi, Kengo Sato, Tetsuro Nishino, and Etsuji Tomita, 32–44. Berlin: Springer-Verlag.

de la Higuera, Colin. 2010. *Grammatical inference: Learning automata and grammars*. Cambridge: Cambridge University Press.

Hopcroft, John, Rajeev Motwani, and Jeffrey Ullman. 2001. *Introduction to automata theory, languages, and computation*. Boston: Addison-Wesley.

Horning, James J. 1969. A study of grammatical inference. Doctoral dissertation, Stanford University, Stanford, CA.

Inkelas, Sharon, and Cheryl Zoll. 2005. *Reduplication: Doubling in morphology*. Cambridge: Cambridge University Press.

Jain, Sanjay, Daniel Osherson, James S. Royer, and Arun Sharma. 1999. *Systems that learn: An introduction to learning theory*. 2nd ed. Cambridge, MA: MIT Press.

Jensen, John. 1974. Variables in phonology. *Language* 50:675–686.

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

Johnson, Kent. 2004. Gold's theorem and cognitive science. *Philosophy of Science* 71:571–592.

Jurafsky, Daniel, and James Martin. 2000. *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Upper Saddle River, NJ: Prentice-Hall.

Jusczyk, Peter, Angela Friederici, Jeanine Wessels, Vigdis Svenkerund, and Ann Marie Jusczyk. 1993. Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* 32:402–420.

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

Karttunen, Lauri. 1998. The proper treatment of optimality in computational phonology. In *FSMNLP'98: Proceedings of the International Workshop on Finite State Methods in Natural Language Processing*, ed. by Lauri Karttunen and Kemal Oflazer, 1–12. Ankara, Turkey: Bilkent University.

Kasprzik, Anna, and Timo Kötzing. 2010. String extension learning using lattices. Paper presented at the 4th International Conference on Language and Automata Theory and Applications (LATA 2010), Trier, Germany, 24–28 May.

Kisseberth, Charles. 1970. On the functional unity of phonological rules. *Linguistic Inquiry* 1:291–306.

Kobele, Gregory. 2006. Generating copies: An investigation into structural identity in language and grammar. Doctoral dissertation, UCLA, Los Angeles, CA.

Kornai, András. 2007. *Mathematical linguistics.* London: Springer-Verlag.

Koskenniemi, Kimmo. 1983. Two-level morphology. Publication no. 11. Helsinki: University of Helsinki, Department of General Linguistics.

Kracht, Marcus. 2003. *The mathematics of language.* Berlin: Mouton de Gruyter.

Manning, Christopher, and Hinrich Schütze. 1999. *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Mattys, Sven, and Peter Jusczyk. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition* 78:91–121.

Mattys, Sven, Peter Jusczyk, Paul Luce, and James Morgan. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology* 38:465–494.

McCarthy, John. 1979. Formal problems in Semitic phonology and morphology. Doctoral dissertation, MIT, Cambridge, MA.

McCarthy, John. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17:207–263.
McNaughton, Robert, and Seymour Papert. 1971. *Counter-free automata.* Cambridge, MA: MIT Press.
Mester, Armin. 1988. *Studies in tier structure.* New York: Garland.
Miller, Phillip H. 1999. *Strong generative capacity: The semantics of linguistic formalism.* Stanford, CA: CSLI Publications.
Mpiranya, Fidèle, and Rachel Walker. 2005. Sibilant harmony in Kinyarwanda and coronal opacity. Talk handout, GLOW 28, University of Geneva.
Myers, Scott. 2002. Gaps in factorial typology: The case of voicing in consonant clusters. Ms., University of Texas at Austin. Rutgers Optimality Archive ROA 509-0302. http://roa.rutgers.edu.
Nevins, Andrew. 2005. Conditions on (dis)harmony. Doctoral dissertation, MIT, Cambridge, MA.
Ní Chiosáin, Maire, and Jaye Padgett. 2001. Markedness, segment realization, and locality in spreading. In *Constraints and representations: Segmental phonology in Optimality Theory*, ed. by Linda Lombardi, 118–156. Cambridge: Cambridge University Press.
Niyogi, Partha. 2006. *The computational nature of language learning and evolution.* Cambridge, MA: MIT Press.
Nowak, Martin A., Natalia L. Komarova, and Partha Niyogi. 2002. Computational and evolutionary aspects of language. *Nature* 417:611–617.
Odden, David. 1994. Adjacency parameters in phonology. *Language* 70:289–330.
Onishi, Kristine H., Kyle E. Chambers, and Cynthia Fisher. 2003. Infants learn phonotactic regularities from brief auditory experience. *Cognition* 87:B69–B77.
Onn, Farid. 1980. Aspects of Malay phonology and morphology. Doctoral dissertation, Universiti Kebangsaan Malaysia, Bangi.
Osherson, Daniel, Scott Weinstein, and Michael Stob. 1986. *Systems that learn.* Cambridge, MA: MIT Press.
Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33:999–1035.
Pater, Joe, and Anne-Michelle Tessier. 2006. L1 phonotactic knowledge and the L2 acquisition of alternations. In *Inquiries in linguistic development: Studies in honor of Lydia White*, ed. by Roumyana Slabakova, Silvina A. Montrul, and Philippe Prévost, 115–131. Amsterdam: John Benjamins.
Piattelli-Palmarini, Massimo, ed. 1980. *Language and learning: The debate between Jean Piaget and Noam Chomsky.* Cambridge, MA: Harvard University Press.
Poser, William. 1982. Phonological representation and action-at-a-distance. In *The structure of phonological representations*, ed. by Harry van der Hulst and Norval Smith, 121–158. Dordrecht: Foris.
Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical Report 2, Rutgers University Center for Cognitive Science, New Brunswick, NJ.
Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar.* Malden, MA: Blackwell.
Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Fixing priorities: Constraints in phonological acquisition*, ed. by René Kager, Joe Pater, and Wim Zonneveld, 245–291. Cambridge: Cambridge University Press.
Pulleyblank, Douglas. 2002. Harmony drivers: No disagreement allowed. In *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, ed. by Julie Larson and Mary Paster, 249–297. Berkeley: University of California, Berkeley Linguistics Society.
Raimy, Eric. 2000. *The phonology and morphology of reduplication.* Berlin: Mouton de Gruyter.
Reiss, Charles. 2009. Long-distance dependencies and other formal issues in phonology. In *Contemporary views on architecture and representations in phonology*, ed. by Eric Raimy and Charles Cairns, 247–257. Cambridge, MA: MIT Press.
Reiss, Charles, and Frédéric Mailhot. 2007. Computing long-distance dependencies in vowel harmony. *Biolinguistics* 1:28–48.
Riggle, Jason. 2004. Generation, recognition, and learning in finite state Optimality Theory. Doctoral dissertation, UCLA, Los Angeles, CA.

Roark, Brian, and Richard Sproat. 2007. *Computational approaches to morphology and syntax.* Oxford: Oxford University Press.
Rogers, James, Jeffrey Heinz, Matt Edlefsen, Dylan Leeman, Nathan Myers, Nathaniel Smith, Molly Visscher, and David Wellcome. To appear. On languages piecewise testable in the strict sense. In *Proceedings of the 11th Meeting of the Association for Mathematics of Language.*
Rogers, James, and Geoffrey K. Pullum. To appear. Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information.*
Rose, Sharon, and Rachel Walker. 2004. A typology of consonant agreement as correspondence. *Language* 80:475–531.
Sagey, Elizabeth. 1986. The representation of features and relations in non-linear phonology. Doctoral dissertation, MIT, Cambridge, MA.
Samuels, Bridget. 2009. The structure of phonological theory. Doctoral dissertation, Harvard University, Cambridge, MA.
Sapir, Edward, and Harry Hoijer. 1967. *The phonology and morphology of the Navaho language.* Berkeley: University of California Press.
Schein, Barry, and Donca Steriade. 1986. On geminates. *Linguistic Inquiry* 17:691–744.
Schoonbaert, Sofie, and Jonathan Grainger. 2004. Letter position coding in printed word perception: Effects of repeated and transposed letters. *Language and Cognitive Processes* 19:333–367.
Shawe-Taylor, John, and Nello Cristianini. 2005. *Kernel methods for pattern analysis.* Cambridge: Cambridge University Press.
Shieber, Stuart. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy* 8:333–343.
Simon, Imre. 1975. Piecewise testable events. In *Automata theory and formal languages*, ed. by Gerhard Goos and Juris Hartmanis, 214–222. Berlin: Springer-Verlag.
Suzuki, Keiichiro. 1998. A typological investigation of dissimilation. Doctoral dissertation, University of Arizona, Tucson.
Tesar, Bruce. To appear. Learning phonological grammars for output-driven maps. In *Proceedings of NELS 39.*
Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory.* Cambridge, MA: MIT Press.
Vapnik, Vladimir. 1998. *Statistical learning theory.* New York: Wiley.
Viljoen, Johannes Jurgens. 1973. *Manual for Ndonga.* Part 1. Pretoria: University of South Africa.
Walker, Rachel. 2000. Long-distance consonantal identity effects. In *WCCFL 19: Proceedings of the 19th West Coast Conference on Formal Linguistics*, ed. by Roger Billerey and Brook Danielle Lillehaugen, 532–545. Somerville, MA: Cascadilla Press.
Walker, Rachel. 2007. Phonetics and phonology of coronal harmony. The case of Kinyarwanda. Talk handout, UCLA Colloquium Series. UCLA, Los Angeles, CA.
Whitney, Carol. 2001. How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin Review* 8:221–243.
Whitney, Carol, and Rita Sloan Berndt. 1999. A new model of letter string encoding: Simulating right neglect dyslexia. *Progress in Brain Research* 121:143–163.
Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30:945–982.

Department of Linguistics and Cognitive Science
University of Delaware
Newark, DE 19716

heinz@udel.edu