

# Computational Phonology – Part II: Grammars, Learning, and the Future

Jeffrey Heinz

December 17, 2010

## Abstract

Computational phonology studies sound patterns in the world's languages from a computational perspective. This article shows that the similarities between different generative theories outweigh the differences, and discusses stochastic grammars and learning models within phonology from a computational perspective. Also, it shows how the hypothesis that all sound patterns are subregular can be investigated, pointing the direction for future research. Taken together, these contributions show computational phonology is identifying stronger and stronger universal properties of phonological patterns, which are reflected in the grammatical formalisms phonologists employ.

This article is intended primarily for phonologists who are curious about computational phonology, but do not have a rigorous background in mathematics or computation. However, it is also informative for readers with a background in computation and the basics of phonology, and who are curious about what computational analysis offers phonological theory.

## 1 Introduction

In *Computational Phonology – Part I: Foundations*, the foundations of the theory of computation were presented and the first major result of computational analysis of phonological patterns was given: they are regular patterns. This article explains why

1. different grammatical formalisms in phonology (in particular, SPE-style phonology, Two-

Level Phonology, Declarative Phonology, and Optimality Theory) are similar to each other from a computational perspective,

2. computational analysis informs stochastic approaches to grammars,
3. computational analysis is of paramount importance in addressing the learning problem, and
4. why the future of computational analysis of phonological patterns is subregular.

Part I emphasized that phonological theories factor the phonology of a language into individual generalizations as shown in Figure 1. It was explained that for SPE-style grammars, Kaplan and

$$F_1 \times F_2 \times \dots F_n = P_L$$

Figure 1: Theories of phonology.

Kay (1994) showed that each rule ( $F_i$ ) encodes a regular relation and that the interaction of these rules ( $\times$ ) is exactly given by the composition operation ( $\circ$ ) of finite-state transducers. Assuming that phonological rules cannot re-apply to the locus of their structural change, it follows that the phonology of the whole language ( $P_L$ ) is also a regular relation, describable with a single finite-state transducer.

In this article, Sections 2-4 show that research on Two-Level Phonology, Declarative Phonology, and Optimality Theory (OT) reveals that the grammars in these theories also describe generalizations and whole phonologies as regular relations, which interact in similar ways. Sections 5-7 discuss stochastic grammars, learning, and subregular properties, respectively. Section 8 concludes.

## 2 Two-level Phonology

**Architecture.** Inspired by Kaplan and Kay (1981), Koskenniemi (1983) presents a theory of phonology called TWO-LEVEL PHONOLOGY (2LP).<sup>1</sup> Under this theory, all individual phonolog-

---

<sup>1</sup>Koskenniemi called his theory “Two-Level Morphology” since he describes morphophonological alternations.

ical generalizations ( $F_i$ ) are stipulated to be regular relations, but these relations are no longer conceived as rules that transform one string to another. Rather, they are individual constraints that enforce well-formedness conditions on alternations in parallel. These conceptual shifts ought not be underestimated as they foreshadow the later development of OT (Prince and Smolensky, 2004).

The interaction of these constraints is accomplished through what I will call the **PATH PRODUCT** operation on finite-state transducers. Examples of paths are given in Part I. Every transducer not only describes a regular relation, it also describes a set of paths. The path product of  $T_1$  and  $T_2$  is an operation that yields another transducer which describes exactly the intersection of the paths of  $T_1$  and  $T_2$ . Kaplan and Kay (1994) explain this operation in more detail.

Karttunen (1993) makes clear that the 2LP grammars decompose the phonology of the whole language in a different manner than the way SPE grammars do. The decomposition in 2LP is just as logically consistent as the one in SPE. Both state individual phonological generalizations as regular relations. 2LP uses path product, but SPE uses composition.

**Expressivity.** Kaplan and Kay (1994) apply their analysis to two-level phonology and show that it, like the SPE-style grammars, expresses all and only regular relations. They conclude “Although there may be aesthetic or explanatory differences between the two formal systems, empirical coverage by itself cannot be used to choose between them.”

**Computational Complexity.** Barton *et al.* (1987) determine the generation problem (part I, problem 4-a) for 2LP grammars belongs to NP, when the grammars are given as a set of individual generalizations. In other words, there is no deterministic algorithm that is guaranteed to compute the surface form from any underlying form for any 2LP grammar in fewer than  $f(n)$  steps where  $f$  is a polynomial function and  $n$  is the length of the underlying form. This is due to the fact that taking the path product of arbitrarily many FSMs belongs to NP (Hopcroft *et al.*, 2001).

Barton *et al.* (1987) make clear that complexity results of this kind ought to lead to future work. Complexity results necessarily consider the worst case, but whether actual phonologies are

among the worst cases remains unknown. Given the earlier observation that phonological patterns are almost certainly subregular, it is a distinct possibility that actual phonologies are not among the worst cases, but instead belong to less expressive, but more computationally tractable, areas of the regular languages.

### 3 Declarative Phonology

**Architecture.** Declarative Phonology (DP) (Scobbie, 1993; Coleman, 2005), like 2LP, maintains that each phonological generalization ( $F_i$ ) is an exceptionless constraint expressed as a regular relation and every constraint must be satisfied by every form. But the interaction of these relations is via set intersection ( $\cap$ ). Consequently, the whole phonology maps an underlying form  $u$  to a surface form  $s$  if and only if every individual phonological generalization maps  $u$  to  $s$ . In FSM terms, set intersection ( $\cap$ ) is equivalent to the automata product ( $\times$ ) operation, which is distinct from the path product.

One difference between DP and 2LP is that a lexical item is explicitly given as a finite language which contain all and only its surface variants. For example, consider a prefix which ends with a nasal, which always agree in place with the following consonant (and assume there are no vowel-initial words). As it is impossible to ascertain the place of the nasal, the prefix is represented as a finite state acceptor which describes the language  $\{im-, in-, ij-\}$ . This is similar to underspecification analyses in phonology (Archangeli, 1988; Mohanan, 1991; Ito *et al.*, 1995; Steriade, 1995) (see also Hooper (1976)).

Another difference between DP and other theories is that the individual generalizations were stated with logical formulae because DP's proponents wanted to emphasize the denotational nature of phonological constraints with declarative statements, independent of any particular implementation.

**Expressivity.** Regular relations are not closed under intersection. This means that the intersection of two regular relations may not be regular.<sup>2</sup> Thus even though DP presumes the constraints are regular, there is nothing in the theory that ensures that the phonology of the whole language ( $P_L$ ) is regular. Thus, DP is strictly more expressive than either SPE or 2LP and consequently, as a scientific hypothesis about what is a possible phonological pattern, is a weaker theory than either SPE or 2LP. As far as I know, this additional expressivity is also unnecessary as no bonafide context-free phonological pattern has been established.

**Computational Complexity.** Computing the product of arbitrarily many FSMs belongs to NP (Hopcroft *et al.*, 2001). Consequently, as with 2LP, it remains to be learned whether further restrictions on possible constraint in DP alleviates this intractability.

## 4 Optimality-Theoretic Phonology

**Architecture.** SPE, 2LP, and DP view the individual phonological generalizations ( $F_i$  in Figure 1) as language-specific. On the other hand, in OT every language has the same individual phonological generalizations (Prince and Smolensky, 1993; McCarthy and Prince, 1995; Prince and Smolensky, 2004), and languages only differ in how those generalizations interact. Like DP and 2LP, the individual generalizations are constraints. But the interaction ( $\times$ ) of constraints in OT is more subtle than in either of those theories: the constraints are prioritized and underlying forms are mapped to surface forms which optimally satisfy these constraints. Standardly, constraints are stipulated to come in two kinds: MARKEDNESS CONSTRAINTS, which penalize alternations with marked surface forms, and FAITHFULNESS CONSTRAINTS, which penalize surface forms which deviate from the underlying form (Kager, 1999; Prince and Smolensky, 2004; McCarthy, 2007b).

Formally, the architecture of OT grammars admits three main components. GEN relates underlying forms to a potentially infinite set of candidate surface forms. CON is a list of totally ordered

---

<sup>2</sup>Beesley and Karttunen (2003) explain clearly why on pages 54-56.

constraints, each of which maps an underlying form  $u$  and a candidate surface form  $s$  to a non-negative integer (the number of times  $(u, s)$  violates the constraint). Thus as a whole, CON maps each  $(u, s)$  pair to a vector of numbers called the VIOLATION PROFILE. EVAL selects, for a given underlying form, the candidate surface form with the most favorable violation profile: the one that violates the most important constraints the least.

Prince (2002) establishes the logic of constraint interaction and procedures for reasoning about OT grammars from the essential information provided by violation profiles. This work makes clear the logical foundation for the inference of constraint ranking from partial information originally explored by Tesar (1995), which is discussed in more detail below.

**Expressivity.** How restrictive is the theory? In a departure from 2LP and DP, the founders of OT made no assumption that it ought to “be regular.” As Eisner (1997c) observes, without restricting what counts as a possible markedness or faithful constraints, there are hardly any limits on the expressivity OT. For example, it is generally assumed that a context-free markedness constraint which assigns violations to words with with a string of  $n$  consonants followed by  $n$  vowels ( $C^nV^n$ ) does not belong to the CON, but nothing in the theory precludes this.

Such constraints are probably tacitly omitted because of the intuitions phonologists have that no phonology permits such an individual phonological generalization, which would be manifest in some language and therefore desirable to include in the factorial typology. Indeed, much phonological research examines the expressivity of OT grammars by comparing predicted typological facts to known typological facts. Idsardi (1998, 2000) and Bakovic (2005) establish that OT is unable to describe attested opaque alternations. Other research establishes that the interaction of plausible constraints either neatly matches the relevant typology (Prince and Smolensky, 1993; McCarthy and Prince, 1995; Kager, 1999) or leads to undesirable typological predictions due to the unexpected ways constraints interact (Eisner, 1997a,c; Wilson, 2001; McCarthy, 2003; Gordon, 2007). Antilla (2008) describes a procedure which calculates a TYPOLOGICAL ORDER from a given set of candidates with their violation profiles, which provides another way to measure the predictions a set of

constraints makes against the known typology.

Given earlier results that showed phonological patterns are regular, other researchers have asked what conditions need to be placed on CON and GEN to ensure that an OT grammar describes a regular relation. Frank and Satta (1998) show that if the only requirements are that GEN and the contents of CON be regular then OT grammars can describe context-free relations. Placing an upper bound on the number of violations regular constraints in CON can assign, however, reduces the expressivity of OT grammars to regular relations. Like Kaplan and Kay's work, their proof is constructive. Karttunen (1998) achieves the same conclusion constructively and further expresses the interaction of the constraints precisely in terms of an operation called LENIENT COMPOSITION (because it is defined in terms of ordinary composition and a special union operation over relations). Together these researchers show that, under the assumptions stated above, that OT grammars combine regular OT constraints in a manner which describes a single regular relation. The importance of this result is that differences between OT grammars and SPE-style rule grammars fade away: like 2LP and DP, they are both particular compositions of regular relations which ultimately describe *the same* regular, functional characterizations of phonological patterns. This point is made repeatedly by Karttunen (1993, 1998).

Another question is whether every regular relation can be expressed with an OT grammar. This question has not been answered. I venture that, unlike Kaplan and Kay's (1994) work, it is not the case that any regular relation can be described by an OT grammar, at least with traditional (and regular) markedness and faithfulness constraints. This is because there are regular relations which describe opaque alternations (because they are describable with SPE-style phonologies), which cannot be described by OT grammars (McCarthy, 2007a).

**The Generation Problem.** Ellison (1994) was the first to give a solution to the generation problem (Part I, 4-a) in OT. Frank and Satta (1998) provide a simpler solution, and Karttunen's (1998) lenient composition operator makes their solution clearer. Gerdemann and van Noord (2000) improve Karttunen's result in the sense that they obtain a more efficient implementation.

Riggle's (2004) constructive, computational analysis of OT is especially noteworthy because he shows that the generation problem in OT can be solved using Dijkstra's shortest paths algorithm (Dijkstra, 1959). Riggle presents a ranking-independent representation of EVAL as a finite-state transducer where the transitions are 'weighted' with violation profiles. This FSM is constructed through a special product operation over the constraints in CON (represented as transducers) called M-INTERSECTION. Given the ranking of constraints and this representation, Dijkstra's algorithm finds the optimal path through the machine in linear time. Furthermore, unlike earlier approaches, there is no need to place an upper bound on the number of the number of violations constraints can assign to achieve this result.

Riggle also shows how to efficiently generate the CONTENDERS—all and only those candidates that could win under any possible ranking, which provides a solution to the typological problem (Part I, 4-d). Even though Riggle's GEN produces infinitely many candidates, in a variety of simulations, the contenders are few.

Recalling Frank and Satta's results, an open question remains: Is the particular GEN function which Riggle employs one that permits the construction of a regular transducer for any (regular) CON and ranking? More generally, what restrictions are necessary on EVAL, GEN or CON to guarantee that optimization yields a regular relation? Riggle (2004b, Chap. 7) presents some preliminary, promising work addressing this issue.

**Computational Complexity.** Eisner (1997b, 2000) building on work by Ellison (1994), was the first to establish that the generation problem in OT is NP as the number of tiers or constraints grows arbitrarily large. Wareham (1998) proves the same result with a different proof and Idsardi (2006) presents an alternative version of Eisner's proof using phonologically standard constraints. Heinz *et al.* (2009) qualify their results by identifying the source of hardness to be computing the product of individual constraints. (Recall that the automata product of individual FSMs, which represent constraints, belongs to NP.) If the constraint set is represented differently (with Riggle's (2004) ranking-independent EVAL) then, as mentioned, the generation problem becomes linear.

But if EVAL is represented in terms of individual constraints then their compilation into a ranking-independent EVAL with M-intersection is also NP.

**OT Variants.** There are many distinct variants of OT; three are discussed here. The original version (Prince and Smolensky, 1993), dubbed CONTAINMENT THEORY, requires that all candidates produced by GEN to be at least as long as the underlying form, with special markers indicating deletions and epenthesis. With the exception of Riggle (2004a), it is this version that has been subject to the computational analyses above.

McCarthy and Prince (1995) introduce CORRESPONDENCE THEORY where underlying forms are indexed and GEN produces every logical possible indexing over every logically possible word. Thus the underlying form  $/k_1 \text{æ}_2 t_3/$  has fully merged candidates  $[k_{1,2,3}]$ , partially merged candidates  $[k_{1,2}]$ , ‘reverse’ candidates like  $[k_3 \text{æ}_2 t_1]$  and  $[t_3 \text{æ}_2 k_1]$ , epenthetic candidates  $[k_1 \text{æ}_2 t_3 s]$ , split candidates  $[k_1 \text{æ}_2 t_3 s_1]$ , and so on.

McCarthy (2008) presents HARMONIC SERIALISM (HS), where GEN is very restricted. Informally, for each underlying form, the candidates produced by GEN can differ by at most one single change, where a single change is roughly speaking, a single insertion, deletion, or substitution. If this was all there was to it, then Harmonic Serialism would be inadequately expressive since many phonological alternations exhibit several changes between underlying forms and their surface forms. However, HS asserts that EVAL and GEN are called repeatedly: for an underlying form  $u$ , EVAL selects the best candidate  $c_1$  from  $Gen(u)$ , and then EVAL selects the best candidate  $c_2$  from  $Gen(c_1)$  and so on until candidate  $c_{n+1} = c_n$ . In this way HS is a model which is both derivational and constraint-based. Taking what is meant by “single change” at face value, it appears this GEN is regular, which would suggest that the kinds of relations HS describes are subject to the results of Frank and Satta’s, Karttunen’s and Riggle’s analyses.

**Summary.** Table 1 summarizes the architectures of the different formalisms and some of the known and unknown facts. Because finite-state transducers, which can describe all and only the

	$F_i$	$\times$	other conditions	properties of $P_L$	
				necessary	sufficient
SPE	regular, lg. specific	composition	rules cannot re-apply to locus of structural change	is regular	is regular
2LP	regular, lg. specific	path product	none	is regular	is regular
DP	regular, lg. specific	intersection	none	is context-free	?
OT-CT	regular, universal	lenient composition	none	is context-free	?
			upper bound on violation profile	is regular	?
OT-R04	regular, universal	M-intersection	none	?	?

Table 1: The architectures of different phonological theories expressing how different factors (i.e phonological generalizations) interact to describe the phonology of a language, recalling that  $F_1 \times F_2 \times \dots \times F_n = P_L$ . OT-CT and OT-R04 are the containment theory and Riggle’s (2004) variants, respectively.

regular relations, are solutions to both the generation and recognition problems (Part I, 4-ab), the theories which necessarily describe the whole phonology as a regular relation, have solutions to both of these problems.

## 5 Stochastic Phonology

Recall the optional rule of word-final coronal stop deletion in English from Part I, repeated below.

$$(1) \quad [+coronal, -continuant] \rightarrow \emptyset / C \_\_\_\_\_\_ \#$$

Functional characterizations of phonological generalizations with a codomain of  $\{0,1\}$  as shown in Part I are unable to distinguish the frequency with which the rule applies, and are unable to make

finer distinctions in the grammaticality or acceptability of possible surface variants.

For these reasons, sociolinguists and phonologists make use of stochastic grammars, which not only describe functional characterizations with real codomains, but also leverage results from statistics and probability theory.

Although it may be surprising, it is nevertheless true: changing the codomain of the functional characterizations from  $\{0,1\}$  to the set of real numbers in this way does not change their expressivity. A regular language which is made stochastic is still regular (Vidal *et al.*, 2005a,b). In fact, every region of the Chomsky Hierarchy admits a probabilistic counterpart (Booth, 1969; Charniak, 1996). Thus although in some sense probabilistic grammars allow finer distinctions to be made, in an important sense they are fundamentally the same as their non-probabilistic counterparts. This is because formal language theory determines expressivity based on structural properties of grammars, which are independent of whether the codomain of the functional characterization of the language they describe is real or boolean.

**Free variation.** Within SPE-style grammars, each rule can be assigned a probability. The optional rule in (1) can be given as follows.

(2)  $[+coronal,-continuant] \rightarrow \emptyset / C \_\_\_\_\_\_ \#$  (0.3)

This means at the stage of the derivation when the rule could apply, it applies 30% of the time (Sankoff, 1978; Sankoff and Labov, 1979; Guy, 1980). In terms of the functional characterization of this generalization, pairs of forms are mapped to real values. A fragment of the generalization stated in (2) is shown in Figure 2. Anttila (1997) and Boersma (1997) provide different approaches to free variation within OT.

**Gradient acceptability judgements.** Phonologists also employ probabilistic grammars to predict gradient acceptability judgements humans give to novel word forms under laboratory conditions. Although well-formedness and likelihood are logically distinct notions, this research makes the

	$f$	
(wɛst,wɛst)	→	0.7
(wɛst,wɛs)	→	0.3
(wɛst,wɛ)	→	0
	...	
(pɪwɛst,pɪwɛst)	→	0.7
(pɪwɛst,pɪwɛs)	→	0.3
(pɪwɛst,pɪwɛ)	→	0
	...	

Figure 2: Fragments of the functional characterization of the post-consonantal, word-final coronal stop deletion rule which applied 30% of the time.

strong hypothesis that they are in fact the same (Coleman and Pierrehumbert, 1997; Albright and Hayes, 2003; Frisch *et al.*, 2004; Coetzee and Pater, 2007; Coetzee, 2008; Hayes and Wilson, 2008; Antilla, 2008; Albright, 2009; Pater, 2009; Heinz and Rogers, 2010; Heinz and Koirala, 2010).

## 6 Learning Phonology

There are different characterizations of the learning problem (Gold, 1967; Anthony and Biggs, 1992; Jain *et al.*, 1999; Vapnik, 1998; de la Higuera, 2010). Learners, however, are always characterized as functions which map data to grammars. Furthermore, learners are always evaluated in terms of the class of concepts (i.e languages, relations, stochastic languages, etc.) they are guaranteed to learn under defined criteria of what it means to “learn.” This makes sense from the computational perspective discussed in Part I. From the computational perspective, learning algorithms are not just step-by-step procedures, they are step-by-step procedures that solve well-articulated problems. For learning algorithms proposed within phonology, we ought to ask What problem does this algorithm solve?

Recursive Constraint Demotion (RCD) is an algorithm in exactly this sense. Essentially, RCD solves the alternation learning problem (Part I, 4-e). Consider any finite set of computable constraints CON. Let a data point be an underlying form paired with its surface form and consider any finite set of data points. Question: What rankings of CON (if any) are consistent with the set of

data points? RCD provably answers this question correctly in every case (Tesar, 1995; Tesar and Smolensky, 1998, 2000).<sup>3</sup>

On the other hand, some of the proposed variants of RCD (Tesar, 1999; Hayes, 2004; Prince and Tesar, 2004) are not algorithms in this sense. They are clear step-by-step procedures, but it is not known what problem they solve, in the computational sense of the word. Instead these variants are motivated in response to claims that RCD is inadequate in certain respects, and simulations are run in order to get a sense of whether the proposed variant appears to resolve this inadequacy. Often the simulation leads to observations that in some cases the inadequacy is resolved, but in others it is not. The next step in this process is to characterize these cases, and then to modify the variant further. If the simulations appear to work in every case, then it may be possible to state and prove which problem the variant is actually solving.

In my opinion, the simulation-based methodology is not the best way to develop an algorithm which provably solves a particular problem (RCD being a case-in-point). This is not to say that simulation-based research is not addressing real inadequacies with existing algorithms or that the ideas present in such procedures fail to capture important intuitions. However, I would argue that there is no replacement for a procedure which provably solves a problem. Knowing what problem the algorithm solves means intimately understanding the algorithm's behavior. In this case, running the algorithm is no longer a simulation; it is a *demonstration*. In the growing field of modeling the acquisition of sound patterns, it is not a reach to say that simulations are the norm, and demonstrations are a rarity. One only need to witness the role RCD played in the development of OT to understand the power of procedures that are provable solutions to problems.

An instructive example of simulation-based research is from Gildea and Jurafsky (1996). They begin with OSTIA (Oncina *et al.*, 1993), an algorithm that provably solves a problem: this algorithm identifies in the limit from positive data (Gold, 1967) a proper subclass of regular relations, those that are describable by a *subsequential transducer*. Even though Gildea and Jurafsky show that the

---

<sup>3</sup>Technically, RCD returns a particular partial ordering of CON. But minor modifications to RCD allow one to generate all total rankings consistent with the data set (Riggle, 2009).

SPE-style English tapping rule can be represented by a subsequential transducer, they show that OSTIA fails to learn the rule, even when presented with a corpus of ~50,000 pairs of underlying forms and surface forms adapted from the Carnegie Mellon University (CMU) Pronouncing Dictionary, which were uniformly modified to exemplify the rule where it applied. They next modify OSTIA in three ways to capture the intuitions that similar sounds behave similarly, that underlying sounds are similar to the sounds they surface as, and that structural descriptions of rules may contain variables. They show that this modified OSTIA returns a transducer which much more closely represent the English tapping rule than what was returned by the original OSTIA algorithm.<sup>4</sup>

There are three important questions this work leaves us with, which have never been answered satisfactorily. First, why does OSTIA fail to learn the tapping rule? Gildea and Jurafsky hint at the answer: the kind of data that OSTIA would need to see to converge to the tapping rule is not present in the CMU dictionary. Importantly, it is not an issue of quantity, it an issue of kind: OSTIA likely needs to see forms that would never occur in any English lexicon; for example, words with three adjacent [t]s. The second and third questions are What class of rules can the modified OSTIA learn (i.e. What is the problem this algorithm solves?) and Why?

The answers to these questions are not yet known but tools exist to find them. It is often the case that algorithms which solve learning problems are able to characterize exactly the kind of data the algorithm needs to be fed in order for it to succeed. An early example comes from Angluin (1982). The grammatical inference community has provided many more such results (de la Higuera, 2005, 2010). Examples relevant to phonology include Heinz (2010a,b). The utility of such characterizations are many. They can be used to develop artificial language experiments and they can be compared with the forms to which children and infants are actually exposed. In the case of OSTIA above, it ought to be possible to identify the kind of data OSTIA needs to learn the English tapping rule and whether such data is present in the CMU dictionary (all without actually running the learning algorithm!)

---

<sup>4</sup>They also study the case of German word-final stop devoicing.

Additional influential work which approaches phonological learning with simulations include Coleman and Pierrehumbert (1997); Ellison (1992); Boersma and Hayes (2001); Albright and Hayes (2003); Hayes and Wilson (2008); Albright (2009); Goldsmith and Xanthos (2009); Tesar (to appear); Heinz (2009) and Goldsmith and Riggle (to appear). Apart from RCD, work within phonology which presents algorithms which provably solve learning problems include Riggle (2009); Magri (2009); Heinz (2010a,b) and Heinz and Rogers (2010). Magri (2010) also proves that the phonotactic learning problem (Part I, problem 4-d) within OT is not in P. Notably, he suggests research “to find phonologically plausible assumptions on generating function and constraint set that make the problem of the acquisition of phonotactics tractable.”

## **7 Subregular patterns**

Recall the hypothesis presented in Part I that phonology is subregular, restated below:

(3) Hypothesis: Phonology  $\subset$  Regular

There are three reasons to be interested in subregular language classes.

First, although any phonological pattern is regular, “being regular” does not make a pattern a possible phonological one. This is because there are many regular patterns which are not phonological (for example patterns that require words to have a number of consonants evenly divisible by some  $n$ ). Classifying phonological patterns according to known subregular language classes helps us understand what kind of regular patterns they are. In turn, this allows strong, restrictive hypotheses of universal properties of phonological patterns to be formulated and examined.

Second, several negative complexity results mentioned above follow when a parameter of the problem includes *all* regular patterns. Virtually every researcher acknowledges that if stronger, assumptions are made, these negative results may disappear. It is perfectly possible—but unknown—whether restricting the parameters of problems to certain subregular classes eases the problems’ complexity.

Third, there is wide consensus that the learning problem is hampered by hypothesis spaces that are too expressive (Vapnik, 1998; Jain *et al.*, 1999; Sober, 2008; Gallistel and King, 2009). If the right restrictive properties are discovered, it is possible that they may contribute to the learnability of phonological patterns (Heinz, 2007, 2009; Tesar, to appear; Magri, 2010; Heinz, 2010a).

**Subregular hierarchies.** The class of regular languages contains a dual hierarchy of nested regions. There is a LOCAL branch and a PIECEWISE branch. Figure 3 shows the proper inclusion relationships in this hierarchy. Rogers and Pullum (to appear) and Rogers and Hauser (2010) pro-

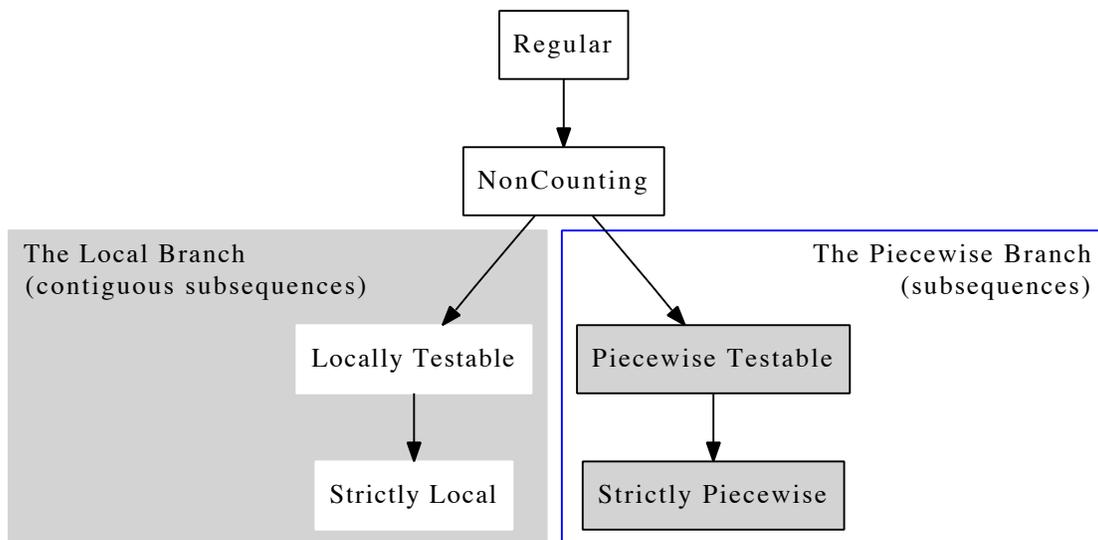


Figure 3: Dual subregular hierarchies (simplified from Rogers *et al.* (2009)).

vides an accessible introduction to the local branch and Rogers *et al.* (2009) to the piecewise branch. The local branch was originally studied by McNaughton and Papert (1971) and the piecewise languages by Simon (1975) and Rogers *et al.* (2009).

Each language class in these hierarchies has independently motivated, converging characterizations and is claimed to correspond to specific, fundamental cognitive capabilities (Rogers and Pullum, to appear; Rogers and Hauser, 2010). Here I provide an informal treatment.

Languages along the local branch only make distinctions on the basis of *contiguous subse-*

quences (up to some length  $k$ , known as  $k$ -factors). For example, the 2-factors of  $\{\times\}abcd\{\times\}$  are  $\{\{\times\}a, ab, bc, cd, d\{\times\}\}$  ( $\{\times\}$  and  $\{\times\}$  are the left and right word boundaries, respectively). Grammars of STRICTLY K-LOCAL ( $SL_k$ ) languages can be thought of as sets of prohibited  $k$ -factors and the words belonging to the language of the grammar are all and only those words which do not contain any of those  $k$ -factors.<sup>5</sup> For example, the grammar  $\{ab\}$  generates the language that contains all words with no contiguous  $ab$  sequence. Grammars of LOCALLY K-TESTABLE ( $LT_k$ ) languages can be thought of as sets of sets of prohibited  $k$ -factors and the words belonging to the language of the grammar are those words whose set of  $k$ -factors is not the same as any of the prohibited sets in the grammar. For example, the  $LT_2$  grammar  $\{\{\{\times\}a, ab, bc, c\{\times\}\}\}$  generates the language that contains all words except those with exactly the set of  $k$ -factors  $\{\times\}a, ab, bc,$  and  $c\{\times\}$ , i.e. all words but  $abc$  (so  $abcd$  is part of the language).

Languages along the piecewise branch only make distinctions on the basis of *subsequences* (up to some length  $k$ , not necessarily contiguous). For example, the 2-long subsequences of  $abcd$  are  $\{ab, ac, ad, bc, bd, cd\}$ .<sup>6</sup> STRICTLY K-PIECEWISE and PIECEWISE K-TESTABLE grammars and languages are defined analogously to ( $SL_k$ ) and ( $LT_k$ ) grammars and languages, respectively.

Finally, the NONCOUNTING class of languages includes all and only those regular patterns which do not count modulo some number  $n$  (so it excludes constraints that penalize words with an even number of consonants, for example). McNaughton and Papert (1971) show that for every noncounting language, there is some  $n$  such that for all logically possible words  $u, v, w$ , either both  $uw^n w$  and  $uv^{n+1}w$  belong to the language or neither do.

**What kinds of subregular patterns are actual phonologies?** Recent work has begun to address where phonological patterns fall in the Subregular hierarchy. A first hypothesis is that markedness constraints are Strictly Local. For example, Hayes and Wilson (2008) use this kind of constraint exclusively, and many markedness constraints in OT appear to belong to SL (though no comprehen-

---

<sup>5</sup>Technically, we let all logically possible words belong to  $\{\{\times\}\} \cdot \Sigma^* \cdot \{\{\times\}\}$ .

<sup>6</sup>Unlike with the Local branch, word boundaries are unnecessary here.

sive survey has been conducted). Locally conjoined markedness constraints (Smolensky, 2006) like  $*ab\&*bc$  are Locally Testable when the domain is the word.<sup>7</sup> Heinz (2010a) hypothesizes that long-distance phonotactic constraints are Strictly 2-Piecewise. Edlefsen *et al.* (2008) examine whether stress patterns are Strictly-Local and if so for what  $k$ . They find most  $\sim 72\%$  are  $SL_k$  for  $k \leq 6$  and that  $\sim 49\%$  are  $SL_k$  for  $k \leq 3$ . Graf (to appear) argues that some stress patterns are not Noncounting.

There are other subregular classes of languages.<sup>8</sup> Heinz (2007, 2009) hypothesizes that all stress patterns and all phonotactic patterns are NEIGHBORHOOD-DISTINCT, an automata-theoretic locality condition, which defines a class which cross-cuts the subregular hierarchies. Heinz (2007) shows that the neighborhood-distinct languages properly includes the  $SL_3$  and  $SP_2$  language classes. No language-theoretic characterization of the neighborhood-distinct class yet exists, though Heinz (2008) begins the analysis.

Tesar (to appear) hypothesizes that the phonological patterns are output-driven. It is not known whether all regular relations can be defined in terms of output-driven maps. I conjecture output-driven regular relations are subregular; i.e. there is at least one regular relation which is not output-driven.

**Subregular relations.** As currently defined, the subregular classes are languages (functions with domain  $\Sigma^*$ ) suitable for describing phonotactics and/or markedness constraints, and not relations (functions with domain  $\Sigma^* \times \Sigma^*$ ), which are suitable for describing alternations. It is an open question how to most fruitfully generalize the subregular hierarchies from sets to relations.

---

<sup>7</sup>The Locally Testable grammar would include all subsets of  $\Sigma^2$  that have both  $ab$  and  $bc$ . If the domain is smaller than the word then the pattern is Strictly  $k$ -Local iff the size of the domain is bounded by  $k$ . Otherwise, the pattern belongs to some to-be-determined subregular class.

<sup>8</sup>Others that have been studied because of their learnability properties include Angluin (1982); Heinz (2008, 2010b)

## 8 Conclusions

The computational study of phonology makes clear that the similarities between competing phonological theories outweigh their differences. In each case, the grammar formalisms decompose the phonology of a language  $P_L$  into individual generalizations, which interact in particular ways. Although the nature of the generalizations and interactions differ in the frameworks, they all have something very important in common: the individual generalizations can all be described as regular relations, their interaction can all be described as a kind of product over regular grammars, and attested phonologies (the  $P_L$  themselves) can be described as regular relations.

Current computational analysis of phonological patterns is revealing stronger, universal properties of phonological patterns. They are subregular. A current hypothesis locates the individual generalizations in the Strictly Local and Strictly Piecewise classes (and their yet-to-be-determined relational counterparts). Stress patterns appear more complex.

It is my hope that this article communicates the importance of the computational perspective to the study of phonology, and more generally to linguistic theory. Several open problems have been identified, which, along with the computational perspective, I hope encourages people to undertake computational analysis when studying the sound patterns in the world's languages.

## References

- Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26:9–41.
- Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90:119–161.
- Angluin, Dana. 1982. Inference of reversible languages. *Journal for the Association of Computing Machinery* 29:741–765.

- Anthony, M., and N. Biggs. 1992. *Computational Learning Theory*. Cambridge University Press.
- Antilla, Arto. 2008. Gradient phonotactics and the complexity hypothesis. *Natural Language and Linguistic Theory* 26:695–729.
- Anttila, Arto. 1997. Variation in Finnish phonology and morphology. Doctoral dissertation, Stanford University.
- Archangeli, Diana. 1988. Aspects of underspecification theory. *Phonology* 5:183–208.
- Bakovic, Eric. 2005. Phonological opacity and counterfactual derivation.
- Barton, G. Edward, Robert Berwick, and Eric Ristad. 1987. *Computational Complexity and Natural Language*. Cambridge, MA: MIT Press.
- Beesley, Kenneth, and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.
- Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences* 21. University of Amsterdam.
- Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32:45–86.
- Booth, Taylor L. 1969. Probabilistic representation of formal languages. In *SWAT '69: Proceedings of the 10th Annual Symposium on Switching and Automata Theory (swat 1969)*, 74–81. Washington, DC, USA: IEEE Computer Society.
- Charniak, Eugene. 1996. *Statistical Language Learning*. MIT Press.
- Coetzee, Andries, and Joe Pater. 2007. Weighted constraints and gradient phonotactics in Muna and Arabic. University of Michigan and University of Massachusetts, Amherst.
- Coetzee, Andries W. 2008. Grammaticality and ungrammaticality in phonology. *Language* 84:218–257.

- Coleman, J. S., and J. 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.
- Coleman, John. 2005. *Introducing Speech and Language Processing*. Cambridge University Press.
- Dijkstra, Esdger W. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik* 1:269–271.
- Edlefsen, Matt, Dylan Leeman, Nathan Myers, Nathaniel Smith, Molly Visscher, and David Wellcome. 2008. Deciding strictly local (SL) languages. In *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics*, edited by Jon Breitenbucher, 66–73.
- Eisner, Jason. 1997a. Efficient generation in Primitive Optimality Theory.
- Eisner, Jason. 1997b. Efficient generation in primitive Optimality Theory. In *Proceedings of the 35th Annual ACL and 8th EACL*, 313–320. Madrid.
- Eisner, Jason. 1997c. What constraints should OT allow? Talk handout, Linguistic Society of America, Chicago. ROA#204-0797. Available at <http://roa.rutgers.edu/>.
- Eisner, Jason. 2000. Easy and hard constraint ranking in Optimality Theory: Algorithms and complexity. In *Finite-State Phonology: Proceedings of the 5th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, edited by Jason Eisner, Lauri Karttunen, and Alain Thériault, 22–33. Luxembourg.
- Ellison, M. T. 1992. The machine learning of phonological structure. Doctoral dissertation, University of Western Australia.
- Ellison, Mark. 1994. Phonological derivation in Optimality Theory. In *COLING 94*, vol. 2, 1007–1013. Kyoto, Japan.

- Frank, Robert, and Giorgio Satta. 1998. Optimality Theory and the generative complexity of constraint violability. *Computational Linguistics* 24:307–315.
- Frisch, S., J. Pierrehumbert, and M. Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22:179–228.
- Gallistel, C.R., and Adam Philip King. 2009. *Memory and the Computational Brain*. Wiley-Blackwell.
- 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.
- Gildea, Daniel, and Daniel Jurafsky. 1996. Learning bias and phonological-rule induction. *Computational Linguistics* 24.
- Gold, E.M. 1967. Language identification in the limit. *Information and Control* 10:447–474.
- 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. 2007. Typology in optimality theory. *Language and Linguistics Compass*, 1:750–769.
- Graf, Thomas. to appear. Comparing incomparable frameworks - a model theoretic approach to phonology. In *Proceedings of the 33rd Penn Linguistics Colloquium*.
- Guy, G.R. 1980. Variation in the group and the individual: The case of final stop deletion. In *Locating Language in Time and Space*, edited by William Labov, 1–36. New York: Academic Press.

- Hayes, Bruce. 2004. Phonological acquisition in optimality theory: the early stages. In *Fixing Priorities: Constraints in Phonological Acquisition*, edited by Rene Kager, Joe Pater, and Wim Zonneveld. 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, University of California, Los Angeles.
- Heinz, Jeffrey. 2008. Left-to-right and right-to-left iterative languages. In *Grammatical Inference: Algorithms and Applications, 9th International Colloquium*, edited by Alexander Clark, François Coste, and Lauren Miclet, vol. 5278 of *Lecture Notes in Computer Science*, 84–97. Springer.
- Heinz, Jeffrey. 2009. On the role of locality in learning stress patterns. *Phonology* 26:303–351.
- Heinz, Jeffrey. 2010a. Learning long-distance phonotactics. *Linguistic Inquiry* 41.
- Heinz, Jeffrey. 2010b. String extension learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden.
- Heinz, Jeffrey, Gregory Koble, and Jason Riggle. 2009. Evaluating the complexity of OT. *Linguistic Inquiry* 40:277–288.
- Heinz, Jeffrey, and Cesar Koirala. 2010. Maximum likelihood estimation of feature-based distributions. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, 28–37. Uppsala, Sweden: Association for Computational Linguistics.
- Heinz, Jeffrey, and James Rogers. 2010. Estimating strictly piecewise distributions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden.

- de la Higuera, Colin. 2005. A bibliographical study of grammatical inference. *Pattern Recognition* 38:1332–1348.
- de la Higuera, Colin. 2010. *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press.
- Hooper, J. 1976. *An Introduction to Natural Generative Phonology*. New York: Academic Press.
- Hopcroft, John, Rajeev Motwani, and Jeffrey Ullman. 2001. *Introduction to Automata Theory, Languages, and Computation*. Boston, MA: Addison-Wesley.
- Idsardi, William. 1998. Tiberian Hebrew spirantization and phonological derivations. *Linguistic Inquiry* 29:37–73.
- Idsardi, William. 2006. A simple proof that Optimality Theory is computationally intractable. *Linguistic Inquiry* 37:271–275.
- Idsardi, William J. 2000. Clarifying opacity. *The Linguistic Review* 17:337–350.
- Ito, Junko, Armin Mester, and Jaye Padgett. 1995. Licensing and underspecification in Optimality Theory. *Linguistic Inquiry* 26:571–614.
- Jain, Sanjay, Daniel Osherson, James S. Royer, and Arun Sharma. 1999. *Systems That Learn: An Introduction to Learning Theory (Learning, Development and Conceptual Change)*. 2nd ed. The MIT Press.
- Kager, René. 1999. *Optimality Theory*. Cambridge University Press.
- Kaplan, Ronald, and Martin Kay. 1981. Phonological rules and finite state transducers. Paper presented at ACL/LSA Conference, New York.
- Kaplan, Ronald, and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20:331–378.

- Karttunen, Lauri. 1993. Finite-state constraints. In *The Last Phonological Rule: Reflections on Constraints and Derivations*, edited by John Goldsmith, 173–194. Chicago: University of Chicago Press.
- Karttunen, Lauri. 1998. The proper treatment of optimality in computational phonology. In *FSMNLP'98*, 1–12. International Workshop on Finite-State Methods in Natural Language Processing, Bilkent University, Ankara, Turkey.
- Koskenniemi, Kimmo. 1983. Two-level morphology. Publication no. 11, Department of General Linguistics. Helsinki: University of Helsinki.
- Magri, Giorgio. 2009. A theory of individual-level predicates based on blind mandatory implicatures. Constraint promotion for Optimality Theory. Doctoral dissertation, MIT.
- Magri, Giorgio. 2010. Complexity of the acquisition of phonotactics in Optimality Theory. In *Proceedings of the Eleventh Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, 19–27.
- McCarthy, John. 2003. OT constraints are categorical. *Phonology* 20:75–138.
- McCarthy, John, and Alan Prince. 1995. Faithfulness and reduplicative identity. In *Papers in Optimality Theory*, edited by Jill Beckman, Laura Walsh Dickey, and Suzanne Urbanczyk, no. 18 in University of Massachusetts Occasional Papers in Linguistics, 249–384.
- McCarthy, John J. 2007a. *Hidden Generalizations: Phonological Opacity in Optimality Theory*. London: Equinox.
- McCarthy, John J. 2007b. What is Optimality Theory? *Language and Linguistics Compass* 10:260–291.
- McCarthy, John J. 2008. The gradual path to cluster simplification. *Phonology* 25:271–319.
- McNaughton, Robert, and Seymour Papert. 1971. *Counter-Free Automata*. MIT Press.

- Mohanan, K. P. 1991. On the bases of radical underspecification. *Natural Language and Linguistic Theory* 9:285–325.
- Oncina, José, Pedro García, and Enrique Vidal. 1993. Learning subsequential transducers for pattern recognition tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15:448–458.
- Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33:999–1035.
- Prince, Alan. 2002. Entailed ranking arguments. *Rutgers Optimality Archive* ROA-500. [roa.rutgers.edu](http://roa.rutgers.edu).
- Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in generative grammar. Tech. Rep. 2, Rutgers University Center for Cognitive Science.
- Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishing.
- Prince, Alan, and Bruce Tesar. 2004. Fixing priorities: constraints in phonological acquisition. In *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge, MA: Cambridge University Press.
- Riggle, Jason. 2004a. Generation, recognition, and learning in finite state Optimality Theory. Doctoral dissertation, University of California, Los Angeles.
- Riggle, Jason. 2004b. Generation, recognition and ranking with compiled OT grammars. Boston. LSA 78.
- Riggle, Jason. 2009. The complexity of ranking hypotheses in optimality theory. *Computational Linguistics* 35:47–59.
- Rogers, James, and Marc Hauser. 2010. The use of formal languages in artificial language learning: a proposal for distinguishing the differences between human and nonhuman animal learners. In

- Recursion and Human Language*, edited by Harry van der Hulst, chap. 12, 213–232. Berlin, Germany: De Gruyter Mouton.
- Rogers, James, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome, and Sean Wibel. 2009. 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 Pullum. to appear. Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information* .
- Sankoff, David. 1978. Probability and linguistic variation. *Synthese* 37:217–238.
- Sankoff, David, and William Labov. 1979. On the uses of variable rules. *Language in Society* 8:189–222.
- Scobbie, James. 1993. Constraint violation and conflict from the perspective of Declarative Phonology. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 38:155–167. Special issue edited by Darlene LaCharite and Carole Paradis.
- Simon, Imre. 1975. Piecewise testable events. In *Automata Theory and Formal Languages*, 214–222.
- Smolensky, Paul. 2006. Optimality in phonology II: Harmonic completeness, local constraint conjunction, and feature domain markedness. In *The Harmonic Mind*, 27–160. Cambridge, MA: MIT Press.
- Sober, Elliot. 2008. *Evidence and Evolution*. Cambridge: Cambridge University Press.
- Steriade, Donca. 1995. Underspecification and markedness. In *Handbook of Phonological Theory*, edited by John Goldsmith, 114–174. Cambridge, Mass.: Blackwell.
- Tesar, Bruce. 1995. Computational Optimality Theory. Doctoral dissertation, University of Colorado at Boulder.

- Tesar, Bruce. 1999. Robust interpretive parsing in metrical stress theory. In *The Proceedings of the West Coast Conference on Formal Linguistics 17*, edited by Kimary N. Shahin, Susan J. Blake, and Eun-Sook Kim. Stanford, CA: CSLI Publications.
- Tesar, Bruce. to appear. Learning phonological grammars for output-driven maps. In *Proceedings of NELS 39*.
- Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.
- Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.
- Vapnik, Vladimir. 1998. *Statistical Learning Theory*. New York: Wiley.
- Vidal, Enrique, Franck Thollard, Colin de la Higuera, Francisco Casacuberta, and Rafael C. Carrasco. 2005a. Probabilistic finite-state machines-part I. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27:1013–1025.
- Vidal, Enrique, Frank Thollard, Colin de la Higuera, Francisco Casacuberta, and Rafael C. Carrasco. 2005b. Probabilistic finite-state machines-part II. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27:1026–1039.
- Wareham, H.T. 1998. Systematic parameterized complexity analysis in computational phonology. Doctoral dissertation, University of Victoria.
- Wilson, Colin. 2001. Consonant cluster neutralisation and targeted constraints. *Phonology* 18:147–197.