Learning and learnability in phonology

Adam Albright, MIT
Bruce Hayes, UCLA

To appear in John Goldsmith, Jason Riggle, and Alan Yu, eds. *Handbook of Phonological Theory*. Blackwell/Wiley.

1. Chapter content

A central scientific problem in phonology is how children rapidly and accurately acquire the intricate structures and patterns seen in the phonology of their native language. The solution to this problem lies in part in the discovery of the right formal theory of phonology, but another crucial element is the development of theories of learning, often in the form of machine-implemented models that attempt to mimic human childrens' ability. This chapter is a survey of work in this area.

2. Defining the problem

Before we can develop a theory of how children learn phonological systems, we must first characterize the knowledge that is to be acquired. Traditionally, phonological analyses have focused on describing the set of attested words, developing rules or constraints that distinguish sequences that occur from those that do not. Although such analyses have proven extremely valuable in developing a set of theoretical tools for capturing phonologically relevant distinctions, it is risky to assume that human learners internalize every pattern that can be described by the theory. Indeed, it is entirely possible that there are systematic patterns that hold true of the lexicon either by sheer accident or because of a series of independent historical changes (Ohala 1981 Listener; Bybee 2001 Phonology; Blevins 2004 Evolutionary; Blevins and Garrett 2004 Metathesis; Yu 2004 Explaining). A theory of human learning should be held accountable for only that knowledge that native speakers can also be shown to have learned. Accordingly, we think it is best to begin by sticking to observables, that is, behaviors and intuitive judgments that reflect phonological knowledge that speakers demonstrably possess. We believe that one of the most powerful demonstrations of phonological knowledge is generalization of the pattern to unknown words. Using this criterion, we find support in the literature for at least three distinct types of phonological knowledge.[1]

Speakers possess **phonotactic knowledge**, meaning that they know, at least tacitly, what constitutes a legal word in their language. Halle (1978 Knowledge) gave an oft-cited example in

[1] A side note: for reasons of space we will have nothing to say about a topic of great importance and relevance; i.e. *phonetic* learning. By this we include the induction learning of phonological categories (features, segments) from waveforms (studied by, e.g. Mielke 2005 Modeling; Lin 2005b Learning features; Maye, Werker and Gerken 2002 Infant sensitivity), language-specific patterns of phonetic realization (e.g. Keating 1985 Universal; Kingston and Diehl 1994 Phonetic), and the vast amount of free variation seen at the phonetic level (as in, for example, coarticulation; Fowler (1981 Coarticulation), Manuel and Krakow (1984 Universal), Smith (1992) Temporal). All phenomena covered here are characterizable at the level of contrasting surface entities.

pointing out that *brick* [bɹɪk] is an existing word of English; *blick* [blɪk] does not exist but in principle could be a word of English, while *bnick [bnɪk] could not. Such claims can be validated not only with observations about the English lexicon, but also by observing loanword adaptation (e.g., *B'nai B'rith* [bəneɪ bɹɪθ], with [ə] inserted in /bn/ but not /bɹ/) and by observing experimentally elicited repetitions and ratings of nonce words. An extensive body of work has gathered phonotactic judgments on a variety of languages, documenting systematic cross-linguistic differences in structures that are deemed acceptable. Thus, the first major task of phonological learning is to determine what is phonotactically legal in the target language.

Speakers of languages also have knowledge of **phonological alternations**. When stems and affixes are combined into words, or words into sentences, their component sounds often change in systematic ways. That speakers often internalize these patterns in their grammars is demonstrated by the substantial literature in "wug testing", starting from Berko (1958 Child's acquisition), illustrating that speakers extend patterns of alternation to nonce stems that they learn in an experimental context. For instance the American English flapping alternation (/t/ → [ɾ] / V ___ V̆) is automatically extended to novel forms; examples can be found in the wug test reported in Albright and Hayes (2003 Rules vs analogy), in which nonce verbs such as *drit* [ˈdɹɪt] were often pronounced with a flap in suffixed forms (*dritting* [ˈdɹɪɾɪŋ]). For other recent demonstrations of generalization of alternations to nonce words, see Zuraw (2000 Pattern exceptions), Albright, Andrade, and Hayes (2001 Segmental environments), Pierrehumbert (2002), Ernestus and Baayen (2003 Predicting the unpredictable), and Zhang, Lai and Turnbull-Sailor (2006 Wug-testing).

Lastly, speakers possess knowledge of **patterns of free variation**; for example, in the idiolects of English that the authors speak, it is a predictable fact about any word containing /æ/ before /m/ or /n/ that it may be realized either as [ɛ̃͡ə̃] or as [æ̃], the latter being preferred in more formal contexts. As the research literature in sociolinguistics demonstrates, such cases could be multiplied indefinitely (for an overviews, see Hay and Drager 2007 Sociophonetics, and Pater and Coetzee, this volume). To some extent, free variation can be considered yet another form of alternation: the same word takes on different forms, but in this case alternation is conditioned by the sociolinguistic context, rather than morphological or phonological context.

With this survey in mind, we can state the scientific problem at hand as follows. The goal is to "reverse-engineer" the human system, constructing a complete model that can acquire phonology exactly as people do. The model must be able learn from positive evidence, with no overt correction of its mistakes. It must learn from real-world utterances, parsing them into their component words and morphemes. It must characterize phonological well-formedness at every level (stems, words, phrases), and it must be able to synthesize novel derived, inflected, and variant forms given suitable information about the form of a stem. Its intuitions of well-formedness must match those of humans exposed to the same data; that is, its behavior under psycholinguistic testing should be closely similar. Lastly, at intermediate phases it should make characteristic errors that match those made during acquisition by human children.

To reach this goal, we need both a theory of phonology (representations, rules/constraints, internal organization of the grammar), and a theory of learning. These components are closely interdependent. No learning theory can make progress unless it is also given a hypothesis space that is adequate to characterize the elements of the learned grammar. Often, learnability researchers assume a great deal of *a priori* knowledge from the learner, such as a universal feature set (as in Chomsky and Halle 1968 SPE) or even a complete universal set of phonological constraints (as in Prince and Smolensky 1993/2004 Optimality Theory; Tesar and Smolensky 2000 Learnability; McCarthy 2001 Thematic). Moreover, the study of learnability often has consequences for the theory of grammar: in a number of cases, theorists have advocated particular principles of grammar precisely because they make phonology learnable where it would not otherwise be. We return to this possibility in section 5.

Our chapter is organized along according to phenomena: phonotactics, then alternations. However, we will also see two cross-classifying themes: the theoretical tools proposed, and particular problems faced by theorists.

3. Learning phonotactics

3.1 Evidence concerning acquisition

Very little is known of the mechanisms by which humans learn the phonotactics of the ambient language. However, one result seems fairly well established: that phonotactic learning is precocious, with considerable progress made well before children can utter words. The evidence for this emerged as new experimental techniques designed to assess the knowledge of infants, such as the head-turn preference procedure (Kemler Nelson et al. 1995 Headturn preference) were applied to phonology, notably by Peter Jusczyk and his colleagues. It emerged that English-learning infants of about nine months listen longer to unfamiliar English words (Jusczyk, Friederici, Wessels, Svenkerud, and Jusczyk (1993 Infants' sound patterns) than to (necessarily) unfamiliar words of a similar but unfamiliar foreign language (Dutch), indicating an ability at this age to identify language solely on the set of sounds involved. Furthermore, Friederici and Wessels (1993 Phonotactic knowledge) have shown that when presented with nonce words that contain legal sounds but in attested vs. unattested combinations (*bref*, *murt* vs. **febr*, **rtum*), Dutch-learning 9-month-olds prefer the attested/well-formed ones. Infants even attend to *gradient* differences of well-formedness, preferring words that contain ordinary, common phoneme sequences over words that contain legal but rare ones (Jusczyk, Luce, and Charles-Luce, 1994 Infants' phonotactic).

Such perception studies have important consequences for the study of phonological learning that we believe are underappreciated. In particular, they show that the tradition of observing and analyzing the spoken output of children, while valuable, may provide at best a very indirect view of what the child has actually learned about the adult language. The imperfect outputs that children later produce are indeed related to adult forms in a systematic, rule-governed way, which we believe appropriately treated as phonological grammar.[2] However, the child's own

[2] See for instance the classic study of Smith (1973 Acquisition), and for a careful overview of more recent work, Demuth's chapter in this volume. For a minority view, claiming a mere physiological basis for child mappings, see Hale and Reiss (1998 Formal and empirical). An issue that we do not address here is whether children's productions are most appropriately modeled as a distinct grammar (Kiparsky and Menn 1977 Acquisition), or with the same grammar that is used for comprehension (Smolensky 1996 Comprehension/Production; Pater 2004 Bridging).

mapping from adult forms to her own surface forms arguably is *not learned at all*, but emerges spontaneously, reflecting the child's efforts to systematically simplify her target outputs to something her still-maturing articulatory apparatus can handle.[3] In sum, current evidence suggests that the learning of the phonological pattern of the adult language is mostly an early and silent process, detectible at most indirectly in the child's own speech.

We also wish to emphasize that, other than the determination that phonotactic acquisition is precocious, very little is known. The infant experiments, ingenious though they are, have largely relied on aggregations of forms, and thus have difficulty in zeroing in on particular phonotactic configurations (though for notable exceptions, see Jusczyk, Smolensky and Allocco 2002 English-learning, and Zamuner, Fikkert and Kerkhoff 2006 Acquisition of voicing). More work will be needed before such studies can "join hands" with theoretical modeling, which as we will see is likewise confined to making very coarse empirical prodictions.

3.2 OT models of phonological learning

Turning to the learning models, we will take as our starting point an influential proposal made within Optimality Theory (Prince and Smolensky 1993/2004 Optimality Theory). The scenario given here was first laid out in Tesar and Smolensky (1993 Learnability) and developed by these authors in a series of works, including Tesar (1995 Computational), Tesar and Smolensky (1998 Learnability-LI) and Tesar and Smolensky (2000 Learnability book).

In this framework, the task of the language learner is to discover a grammar that is consistent with the set of observed forms from the target (adult) language. We suppose that a language learner has access to a representative set of input-output pairs,[4] illustrating the mapping from underlying to surface phonological representations (more on this below). For each input representation, there is exactly one winning output, as well as a set (or a way of computing a set) of losing output candidates. The target grammar is a set of constraints, ranked in such a way that higher-ranked constraints correctly eliminate losing candidates and favor the attested winning output. The constraint set is assumed to be universal and innate (part of Universal Grammar); the task of the learner is simply to discover a ranking of these constraints that is compatible with the data (provided that one exists). The hypothesis space is the set of all k! possible rankings, where k is the number of constraints.

As Tesar and Smolensky point out, the comparison of winning and losing candidates is frequently informative in identifying compatible rankings. Imagine, for instance, a language that permits the marked category of voiced obstruents. We assume for concreteness that the learner

[3] On the other hand, the *delearning* of the child's output system does seem to be data-sensitive. Boersma and Levelt (2000 Gradual constraint) and Curtin and Zuraw (2002 Constraint demotion) suggest that when the child alters her system to produce outputs closer to adult speech, the process is guided by a preference first to master those marked structures that are more common in the ambient language. For discussion of recent investigations into the relation between frequency and order of productive mastery, see Demuth (this volume, section 3.6).

[4] A terminological note: here we use *input* to refer to a representation that is fed into the grammar to derive a surface representation (= an *output*). We designate input forms with slashes (/ba/), and candidate output forms with square brackets ([ba], [pa]). Our use of these terms is common in theoretical phonology, but must be distinguished from *input* as empirically observed learning data (the input to a model, the input to the child).

is equipped with the constraints *VOICED OBSTRUENT ("no voiced obstruents in the output") and IDENT(voice) ("output consonants must not differ from input consonants in voicing".) The system receives the input datum /ba/, and (by some means not discussed here) accesses *[pa] as a loser candidate. The pair of candidates [ba] vs. *[pa] for (assumed) underlying /ba/ is informative with respect to constraint ranking. The constraints IDENT(voice) and IDENT(nasal) are *winner preferrers*, since they assign fewer violations to [ba] than to [pa] or [ma].

(1) Comparison of winning and losing candidates: IDENT(voice), IDENT(nasal) » *VOICED OBSTRUENT

/ba/	IDENT(voice)	IDENT(nasal)	*VOICED OBSTRUENT
☞ a. [ba]			*
* b. [pa]	*!		
* c. [ma]		*!	

The constraint *VOICED OBSTRUENT, on the other hand, is (here) a *loser preferrer*, favoring *[pa]. The basic insight is that an OT grammar will derive the right outputs if, for all such pairs, every loser-preferring constraint is dominated by at least one winner-preferrer.

Tesar and Smolensky propose a ranking algorithm, Recursive Constraint Demotion (RCD), that finds grammars that have this property. RCD assumes that the learner is provided with a set of constraints, a data set consisting of winner/loser pairs generated from a grammar of fully and consistently ranked constraints (i.e., no ties, variable ranking, or errors). The algorithm is simple: it starts with all constraints unranked with respect to each other, and all winner-loser pairs unexplained. At each stage, it demotes all of the constraints that prefer unexplained losing candidates so that they are outranked by constraints that prefer only winners or are neutral. It then checks to see which winner/loser pairs have been successfully explained by virtue of having a winner-preferring constraint ranked above all loser-preferring constraints. Once a pair has been explained, it may be removed from consideration. This reduces the set of unexplained losers and (ideally) also reduces the set of loser-preferring constraints, freeing up some constraints for ranking in the subsequent stage. This process is repeated until no explained pairs remain and all constraints have been ranked into **strata** of constraints that are not crucially ranked with respect to one another. Tesar and Smolensky show that this procedure is guaranteed to find a working grammar, provided that one exists (which it must, by the assumption that the data was generated by such a grammar).

In the example in (1), IDENT(voice) and IDENT(nasal) are winner-preferrers since they both favor the correct output [ba] over an incorrect competitor. Thus, the algorithm places both in the top stratum, demoting *VOICED OBSTRUENT (which prefers losers [pa] and [ma]). With this ranking in place, both competitors are successfully eliminated by high-ranking constraints, and *VOICED OBSTRUENT no longer favors any unexplained losers, so it may be ranked, and the algorithm terminates.

As Tesar and Smolensky point out (2000 Learnability, chap. 3-4) RCD is efficient, since at each step the algorithm is guided by the pattern of constraint violations directly toward the right answer. This can be contrasted with learning procedures in which the search proceeds in a less goal-directed fashion, as in the Triggering Learning Algorithm (Gibson and Wexler 1994

Triggers, Niyogi and Berwick 1996 Language learning; Frank and Kapur 1996 Triggers). Moreover, the algorithm is entirely general; it does not depend in any way on the particular language or set of constraints, but covers any problem that can be reduced to an input-output relation and a suitable constraint set.

The fact that RCD arrives at a compatible ranking reliably and efficiently is a strength of the approach. However, this simple version of the algorithm also relies on quite a few potentially limiting assumptions. First, it simplifies the learning task by assuming that quite a few pieces of the solution are given in advance, including the input-output pairings, the set of losing candidates, and the set of constraints. Second, it requires that the training data be free of errors and variation, making it inappropriate for many realistic learning tasks. Finally, while it is guaranteed to find a grammar that is compatible with the given data, it has no mechanism for deciding among multiple compatible grammars; as we will see below, this often leads to unwanted predictions. We discuss these issues in turn.

First, the finding of the losing rival candidates that RCD needs is a difficult computational problem, particular since the set of potential candidates is infinite. Fortunately, the mathematical apparatus of **finite state machines** has made possible considerable progress here: a finite state machine is a formal object that (provided it includes loops) can represent an infinite class of strings and compute their vector of constraint violations. A variety of work has applied finite state machines to the problem of finding OT candidates (Ellison 1994 Phonological; Eisner 1997 Efficient; Albro 1998 Evaluation, 2005 Studies). Riggle (2004 Generation) has shown that finite state machines are particularly useful in finding the "contender" candidates—i.e., those candidates which could win under at least one constraint ranking, and are thus able to motivate constraint rankings.[5]

Second, the claim that the entire constraint inventory is given to the child in advance is understandably controversial. Given the great complexity of phonology, the idea that the full constraint set could have arisen by natural selection strikes many as implausible. Here, efforts have been made to simplify constraint theory, e.g. by arranging constraints in families (Smith 2004 Making), or by using the language learner's self-explored phonetic knowledge to construct constraints (Hayes 1999 Phonetically; Steriade, in press Phonology of perceptibility).

Third, Recursive Constraint Demotion relies on a comparison of winner-loser candidate pairs: for a given input, one candidate wins, and the other loses, requiring certain constraints to be ranked below others. However, when a single input has more than one possible output a contradiction emerges, since on some occasions one ranking may be needed, while on other occasions the opposite is necessary. It is possible to construct OT grammars that generate free variation by letting certain constraint rankings remain unspecified, and letting their rankings be fixed on an utterance by utterance basis (Reynolds 1994 Variation; Anttila 1997a Variation, 1997b Deriving variation; Nagy and Reynolds 1997 Optimality). However, current convergence proofs rest on an assumption that the data is consistent (i.e., has no variability or errors). Other

[5] This approach is similar in spirit to a proposal by McCarthy (2007) to consider only those candidates that are potentially more harmonic than the fully faithful candidate, an output candidate that faithfully retains all specifications of the input.

approaches to free variation are covered briefly below in section 3.4, and in greater detail in Pater and Coetzee (this volume).

3.3 The subset problem in phonotactic learning

All of the shortcomings just discussed hold for the use of RCD for phonology in general. But the particular problem of phonotactic learning is, in one sense, even harder.

When one derives outputs from inputs (e.g., surface forms from underlying forms), it is possible to limit the problem to a set of choices, and one need only discover the correct choice.[6] But phonotactics is not obviously a matter of deriving outputs from inputs; rather, the intent is to classify all the possible phonological strings as legal or illegal. In the usual instance, the language provides only positive data, informing the learner what is legal, but no negative data to indicate what is ill-formed. It is all too easy for learning algorithms to arrive at grammars that classify the observed data as legal, while failing to classify the illegal forms as such. This is an instantiation of the classic Subset Problem for language learning (see Dell 1981 Learnability; Berwick 1986 Learning; Smith 1999 Positional; Hale and Reiss 2003 Subset; and many others).

The Subset Problem manifests itself in a particular way in standard OT, where the usual approach to phonotactics appeals to the concept of the Rich Base: it is assumed that any phonological representation can be an underlying form, and that what is legal on the surface is simply whatever can be derived, under the phonology, from any UR. As Smolensky (1996), Smith (1999) and others point out, the “tightness” of a phonological grammar will depend on the relative ranking of its Markedness and Faithfulness constraints; in general, the lower Faithfulness is, the fewer forms will be permitted on the surface. Unfortunately, learning based on positive evidence frequently leads RCD to rank Faithfulness constraints high, leading to an insufficiently restrictive analysis.

To see how this may happen, we return to the example from (1) above. In this language, the fact that /ba/ surfaces as [ba] and not as [pa] or [ma] is taken as evidence that IDENT(voice) and IDENT(nasal) are both ranked high. Consider now the predictions of this grammar for a hypothetic input with a voiceless nasal (/m̥a/), in a language that has no voiceless nasals. We assume that the learner comes equipped with a markedness constraint against voiceless nasals. Since by assumption the target language does not actually contain any such sounds, the learner has no reason to demote *VOICELESS NASAL. However, if the learner is restricted to positive examples such as /ba/ → [ba] and no negative examples such as /m̥a/ → [ma], there is also no overt evidence that would compel a ranking of *VOICELESS NASAL over IDENT(voice) or IDENT(nasal). Since the faithfulness constraints never favor a loser during training, the RCD ranks them as high as possible in the grammar, incorrectly predicting that the grammar may at least sometimes (on some occasions, or for some speakers) faithfully produce surface [m̥a]. This prediction is incorrect: studies of loanword adaptation show that when speakers are presented with sounds outside their native language, they typically modify them to conform to native language phonotactics.

[6] This can be done by imposing some large, arbitrary upper length on the members of the candidate set.

(2) Inability to rule out [m̥a]

/m̥a/	*VOICELESS NASAL	IDENT(voice)	IDENT(nasal)	*VOICED OBSTRUENT
☞ a. [m̥a]	*			
☞ b. [ma]		*		
☞ c. [pa]			*	

In the present case, letting the learner start out with Markedness ranked above Faithfulness (*VOICELESS NASAL » IDENT(voice), IDENT(nasal)) would be sufficient, at least to rule out [m̥a] in favor of [ma] or [pa]. Study of the phonotactic subset problem in OT quickly showed that simply starting out learning with Markedness high and Faithfulness low is unlikely to work in general, however: with only positive data available, Faithfulness constraints are still likely to be ranked too high (Ito and Mester 2003 Sources). Hayes (2004 Phonological acquisition), Prince and Tesar (2004 Learning) and Tessier (2006 Biases) propose to amplify RCD with heuristics that are designed actively to keep Faithfulness constraints as low as possible, throughout learning. These proposals express a number of key insights into issues that a learner may face in deciding how restrictive the final grammar should be, but it is currently difficult to evaluate them in detail because we have relatively little empirical data concerning how well human learners solve the subset problem.

3.4 The gradience problem in phonotactic learning

Intuitions of phonotactic well-formedness obtained experimentally are characteristically gradient: hypothetical forms can sound perfect (like, for English, [kɪp]), or completely bad (like [bzɑɹʃk]), or, crucially, intermediate (e.g. [fɹɪlg], [snɔɪks], [pwɪp]).[7] If phonological analysis is to provide a complete account of native speaker intuition, it must characterize such gradient intuitions. (The only clear alternative we are aware of is to let the grammar define a sharp binary distinction and to let analogy to existing forms cover the rest. For evidence against this view, see Hayes and Wilson (2008 Maximum entropy), Albright (in press, Gradient).

An important aspect of gradient phonotactic intuitions is that they are characteristically closely related to the frequencies with which particular sequences occur in the lexicon of the language; see for instance Coleman and Pierrehumbert (1997 Stochastic), Frisch, Large and Pisoni (2000 Perception), Bailey and Hahn (2001 Wordlikeness) and Hay, Pierrehumbert and Beckman (2003 Speech perception). Jusczyk et al. (1994 Phonotactic patterns) found this to be so even in the preferences of infants, who have very little experience with the ambient language. Psycholinguists have long been interested in phonotactic gradience, and in order to measure it

[7] Data gathered as part of the research reported in Albright and Hayes (2003 Rules). Average subject ratings on a scale from 1 (worst) to 7 (best) for these five forms were: [kɪp] 5.84, [bzɑɹʃk] 1.50, [fɹɪlg] 2.68, [snɔɪks] 3.00, [pwɪp] 2.89.

they have characteristically used “quick and dirty” models that can compute predicted gradient phonotactic well-formedness scores on the basis of the frequency of existing forms in the language. Among the most common are *n*-gram models, which involve chopping up the existing lexicon into sequences *n* segments long and estimating the probability (absolute or conditional) of each *n*-gram. (For an introduction to *n*-gram models, see Jurafsky and Martin 2000 Speech, chap. 6). By combining probabilities across all the *n*-grams in a word, a probability value for any novel form can be computed. Such probabilities are usually positively correlated with native speaker judgments gathered experimentally (Vitevitch et al. 1996 Phonotactic; Bailey and Hahn 2001 Wordlikeness).

Although *n*-gram models are attractive as a computational model (Heinz 2007 Inductive Learning) and mimic some aspects of gradient well-formedness intuitions, they are probably not adequate as a theory of how speakers learn and represent gradient patterns, because they fail to characterize the *multidimensionality* of phonotactic patterning. Phonological research indicates that a full model would have to include not just segmental *n*-grams, but a variety of non-local factors. In vowel harmony, vowels some distance from one another must agree in certain of their features; and similar patterns are found for anteriority in coronal sibilants and stops, as well as laryngeal features (MacEachern 1999 Laryngeal) (for an overview, see Hansson 2001 diss., and Rose and Walker, this volume). Prosodic elements, like stress and tone, are also part of phonotactics, and they are characteristically non-local, requiring evaluation over windows larger than a fixed *n* segments. The question of how to integrate these multiple types of conditioning contexts in a *n*-gram model remains, as far as we know, unresolved. Furthermore, *n*-grams stated over segments suffer from insufficient generality, since they fail to incorporate features and natural classes. The practical effect is that in any language, a number of *n*-grams judged well-formed by native speakers would have zero frequency in the lexicon.

A number of researchers studying non-local gradient patterns have proposed models that are more sophisticated than *n*-gram counts over sequences of adjacent segments. These include models that parse and count onsets and rhymes (Coleman and Pierrehumbert 1997 Stochastic; Frisch, Large and Pisoni 2000 Perception; Treiman et al. 2000 English speakers') or track nonlocal similar-segment pairs (Frisch, Pierrehumbert and Broe 2004 Similarity; Coetzee and Pater 2008, Weighted). Although these models perform well at the specific tasks at hand by zeroing in on some particular aspect of phonological structure, they do not represent general purpose learning models of how speakers decide which non-local features to attend to in determining the well-formedness of a sequence. We assume, then, that “quick and dirty” models of the kind just discussed are heuristically useful, since they can represent the gradience seen in lexical patterns and which impinges on gradient intuitions. However, for a full-scale theory of how learners discover phonologically significant gradient patterns, we need we need a way of navigating a rich hypothesis space, including representations in terms of features and natural classes, tier structure, and metrical structure. This suggests that the solution will ultimately require combining insights from phonological theory about the correct representation of linguistic structure, and from machine learning about representing probabilistic information in statistical learning.

3.5 Constraint-based approaches to restrictiveness and gradience

One effort to add gradience to constraint-based theory is a stochastic version of Optimality Theory invented by Boersma (1997 How we learn, 1998 Functional). In standard OT, ranking is a purely relative notion (one constraint categorically outranks another). Boersma (1997 How we learn) proposes a modification that makes ranking gradient by assigning a real number to each constraint, its **ranking value**. The grammar is made to behave stochastically as follows: on each speaking occasion, a random **noise** value, sampled from a Gaussian distribution, is added to each constraint's ranking value. The constraints are sorted according to these perturbed values, and the output is then determined according to the standard evaluation procedure for OT. Whenever conflicting constraints have sufficiently close ranking values, this system will generate multiple outputs, since the stochastic noise will cause the two constraints to switch positions on some occasions. Moreover, since ranking values are continuous, the model can produce outputs with a continuous range of probabilities. Thus at first blush, stochastic OT seems a promising candidate for solving the phonotactic gradience problem.

Stochastic OT has been used in a variety of phonological analyses as the basis for treating free variation in input-to-output mappings (see, e.g., Boersma and Hayes 2001). Moreover, just as with non-stochastic OT, stochastic OT has attracted efforts to construct learning algorithms (Boersma 1997 How we learn; Lin 2005a Learning; Maslova, in press, Stochastic OT; Wilson, ms., Luce choice). The Gradual Learning Algorithm (GLA; Boersma 1997 How we learn) works rather similarly to Recursive Constraint Demotion,[8] only gradiently, and in a number of cases is able to learn grammars that generate free variation. Moreover, the GLA is sensitive to the frequency patterns in the learning data, which as mentioned above are an important source of gradient intuitions in phonology. Unfortunately, unlike RCD, the GLA (at least in its current form) is demonstrably unable to arrive at the correct grammar for certain configurations of constraint violations (Pater 2008b Gradual learning). This drawback has led researchers to seek alternative approaches, some of which constitute more radical departures from the standard assumptions of Optimality Theory.

One important idea from computer science has recently gained attention: the principle of Maximum Likelihood, which states that the grammar to be sought is the one that maximizes the probability of the learning data, given the constraints or other grammatical principles available. This probability is in principle a computable value under any model in which the assessed score of any form is expressed as a probability (i.e. its probability of occurring as a word). The intuitive idea is that if the probability of the observed data is maximized, then the probability of the unobserved data—or more precisely, unobserved data that can be excluded by the constraint system; cf. [blɪk]—is minimized, which corresponds to the ordinary goal of phonotactic analysis.

Jarosz (2006 Rich lexicons, in press 'Restrictiveness') adapts a Maximum Likelihood principle to OT learning. The learner starts with a provisional "pseudo-lexicon" consisting of (or sampled from) the rich base, defining a space of potential underlying forms from which any

[8] Unlike RCD, the GLA employs a "symmetrical" reranking scheme, in which loser-preferring constraints are demoted and winner-preferring constraints are promoted. Boersma (1997 How we learn) argues that this symmetrical strategy is necessary to achieve stasis when variation or errors create conflicting data.

given surface form could be derived. To this is added an innate constraint set (of size *k*) and a body of learning data. The basis of Jarosz's approach is to search all *k*! possible rankings of the constraints, distributing probability among them in a way that maximizes the probability of the learning data. (A very similar proposal can be found in Riggle 2006 Entropy). Moreover, since Jarosz's model assigns a probability distribution over grammars, it is able to assign gradient well-formedness predictions in the form of a probability value for each possible word. This a highly principled solution to the phonotactic learning problem, but it comes at the cost of searching of a truly colossal logical space: the set of *k*! constraint rankings, multiplied by the space of all possible underlying forms. In order to scale up to learning scenarios with more realistic numbers of constraints and underlying forms, a more sophisticated strategy is needed for estimating probability distributions over underlying forms (most likely, making use of non-exhaustive sampling), along with some way of evaluating the likelihood of entire sets of rankings, rather than each of *k*! rankings individually.

Hayes and Wilson (2008 Maximum entropy) propose to abandon OT altogether, adopting instead a stochastic constraint-based framework similar to Harmonic Grammar (Legendre, Miyata and Smolensky 1990 Harmonic; Smolensky and Legendre 2006 Harmonic), in which constraints are weighted rather than given OT rankings. Specifically, Hayes and Wilson employ a Maximum Entropy (log-linear) model to find weights for inductively learned markedness constraints that evaluate the probability of surface strings (as opposed to evaluating them as outputs for some hypothesized input, as in standard OT). A benefit of adopting weighted rather than ranked constraints is that standard search algorithms exist that provably converge on an optimal set of weights; for discussion, see Goldwater and Johnson (2003 Learning), Jäger (2004 Maximum Entropy), Pater, Bhatt and Potts (2007 Optimization), Hayes and Wilson (2008 Maximum Entropy), and Boersma and Pater (2008 ms., Convergence). It is worth noting that in this case, the choice of weighted constraints rather than strictly ranked constraints is motivated almost entirely by convergence properties rather than because an Optimality Theoretic grammar would be inadequate for the task at hand. This appears to be one of the first instances in which considerations of learnability have played a role in motivating architectural decisions in phonological theory.

4. Phonological alternations

The learner can make significant progress on the task of learning surface phonotactics by applying the techniques described above to a set of training data consisting of individual words, or perhaps even a rougher parse of the speech stream into (approximately) word-sized units. However, phonotactic learning is not the only task that learners face: they must at the same time refine their segmentations to determine which words are morphologically complex,[9] and begin to compare related words to discover contextual variation in their pronunciation. For example, a child acquiring Dutch would discover that the word-final [t] in [bɛt] 'bed.SG.' corresponds to [d] in the suffixed form [bɛdən] 'bed.PL.'. Discovering and encoding alternations such as [t] ~ [d] is, logically speaking, a more complex task than learning static phonotactics, since it requires

[9] For some algorithmic approaches to segmentation into words and morphemes, see Harris (1955 Phoneme to morpheme), de Marcken (1996 Unsupervised), Brent and Cartwright (1996 Distributional), Goldsmith (2001 Unsupervised), Baroni (2001 Distributional), and Goldwater (2007 Non-parametric).

comparing morphologically related forms, choosing a basic or underlying form, and learning a grammar that can generate the various surface realizations.

In many cases, prior knowledge of phonotactics could give the learner a leg up in discovering alternations, since as has long been noted, alternations frequently find transparent motivation in phonotactic considerations (Kisseberth 1970 Functional; Sommerstein 1974 Phonotactically). For example, the Dutch voicing alternation seen in [bɛt] ~ [bɛdən] 'bed-SG./PL.' is straightforwardly related to a very general ban on final voiced obstruents in Dutch (*[bɛd]). Optimality Theory provides a straightforward way of relating phonotactic learning with learning of alternations, since an initial phase of phonotactic learning can provide the learner with a crucial component of the analysis (*FINAL VOICED OBSTRUENT >> Faithfulness); all that remains is to learn the relative ranking among faithfulness constraints. For example, a child learning Dutch would need to learn that final voiced obstruents are fixed by devoicing rather than, say, nasalization ([bɛ**n**] ~ [bɛdən]) or vocalization ([bɛ**j**] ~ [bɛdən]). This would follow from the rankings IDENT(nasal), IDENT(consonantal) >> IDENT(voice).

There is reason to believe that children take on the task of learning alternations only after they have made a certain amount of headway on learning phonotactics (Hayes 2004 Phonological acquisition; Prince and Tesar 2004 Learning; Tesar and Prince 2007 Using phonotactic). As discussed above, infants show sensitivity to native language phonotactics at ages as young as 9 months, well before they demonstrate systematic knowledge of words or morphological paradigms. Additional evidence that phonotactic distributions are mastered prior to alternations comes from early child productions. Berko (1958 Child's acquisition) tested the ability of English-learning four year olds to apply voicing and epenthesis alternations in the plural and past tense inflections of novel nouns and verbs: *spow+ed* [spoʊ**d**], *rick+ed* [rɪk**t**], bodd+ed [bad**əd**]. For the most part, children's responses either applied the alternations correctly or omitted the suffix completely; that is, children consistently obey the phonotactics of the adult language (voicing agreement in final obstruent clusters, a ban on identical adjacent consonants), even at a stage when they have not completely mastered the alternations. Furthermore, experimental work with adult speakers has shown that prior phonotactic knowledge (in this case, from the native language) facilitates learning of alternations in an artificial language (Pater and Tessier 2003, 2006).

In the following sections, we briefly review some evidence concerning the acquisition of alternations in children, before turning to proposals for how to model the learning of alternations.

4.1 Evidence concerning acquisition of alternations

Compared with knowledge of surface phonotactics, which can be demonstrated in early infancy (see above), relatively less is known about early knowledge of phonological alternations. By looking at child productions, it is possible to show that at least some alternations are acquired fairly early. For example, Aksu-Koç and Slobin (1985 Turkish) describe a Turkish-learning 15-month old who shows correct mastery of vowel harmony in the accusative suffix ([-a] vs. [-e]). However, wug tests investigating productive mastery of alternations often reveal errors even when children are correctly deploying variants of existing words. It appears that adult-like mastery of many alternations does not emerge until significantly later, with children initially preferring invariant (non-alternating) morphemes. Zamuner, Kerkhoff and Fikkert (2006 Acquisition) and Kerkhoff (2007 Acquisition) have shown that Dutch-learning children have difficulty both recognizing and applying final devoicing in the singulars of novel nouns, while they perform much better on non-alternating items. Kazazis (1969 Possible evidence) presents a case study of one Greek-learning child who at age 4;7 systematically failed to apply the phonotactically regular alternation between [ç] before front vowels ~ [x] elsewhere, resulting in erroneous forms such as *é*[x]*ete* 'have.2PL' instead of adult *é*[ç]*ete*. In both of these cases, there is reason to believe that knowledge of the relevant alternations is acquired eventually; for instance, palatalization alternations are completely predictable and are applied automatically by adult Greek speakers.

Lexically restricted alternations, which frequently have no synchronic phonotactic motivation, appear to pose an even greater difficulty. Clahsen, Aveledo and Roca (2002 Development) show that Spanish-learning children often fail to apply irregular changes such as diphthongization (stressless [e], [o] ~ stressed [jé], [wé]) within verbal paradigms, and Clahsen, Prüfert, Eisenbeiß and Cholin (2002 Strong stems) demonstrate that German learners are likewise reluctant to apply umlaut alternations (e.g. [a] ~ [e]) within present tense verbal paradigms. Similarly, Berko (1958 Child's acquisition) found that English-learning 4 year olds were relatively unlikely to produce voicing alternations in stem-final fricatives in the plural of novel nouns (*heaf* ~ *heaves*) (see also Baker and Derwing 1982 Response; Derwing and Baker 1986 Assessing). This can be contrasted with adult speakers, who do at least sometimes extend lexically restricted alternations in similar experimental settings (Berko 1958 Child's acquisition; Zuraw 2000 Patterned exceptions; Albright, Andrade and Hayes 2001 Segmental environments; Albright and Hayes 2003 Rules vs. analogy, Pierrehumbert 2002 Statistical basis). Thus, it appears that the knowledge of such alternations is acquired much later than knowledge of phonotactics and phonotactically motivated alternations.

The picture that emerges is that the task of learning alternations is a difficult one that requires significant lexical knowledge, and which is taken on gradually over the first 5–10 years of life. Furthermore, although prior knowledge of phonotactics is certainly helpful, it by no means predetermines knowledge of alternations. It appears that even when the relevant phonotactic is known, learners must nonetheless compare related forms and encode alternating variants in some fashion. The procedure that is needed to do this depends intimately on the grammatical mechanism that is employed to encode alternations.

In this section, we review two major approaches, and some challenges.

4.2 Theories for learning alternations I: approaches using underlying forms

One very widely used strategy for encoding alternations is to provide each morpheme with a single unified representation (the underlying form/representation, or UR), and to set up a grammar that derives all observed surface variants from the same underlying form. (Pāṇini; Bloomfield 1933 Language , Chomsky and Halle 1968 SPE). For example, a learner of Dutch confronted with related forms [bɛt] ~ [bɛd-ən] 'bed.SG./PL.' would be forced to select a single underlying form—/bɛt/, /bɛd/, or something more abstract—and derive the surface alternation by intervocalic voicing (/bɛt-ən/ → [bɛdən]) or final devoicing (/bɛd/ → [bɛt]). In the Dutch case, the choice can be made relatively straightforwardly by observing the simultaneous existence of non-alternating voiceless morphemes ([vuːt] ~ [vuːtən] 'foot-SG./PL.'), making intervocalic voicing an untenable solution. In the general case, however, learning a suitable combination of URs + grammar can be difficult because of the circularity involved: the optimal choice for URs depends on having a reasonably good hypothesis about the grammar, but the grammar cannot be formulated without a hypothesis about the set of input → output mappings that it must perform.

The problem of simultaneously learning underlying forms and a grammar that makes use of them is an instance of the more general problem of *hidden structure*: the grammar depends on distinctions that are not part of the immediately observable phonetic context, but rather are structural entities encoded on a language-particular basis. In the Dutch case, the difference between the behavior of the final stops in [bɛt] ~ [bɛdən] and [vuːt] ~ [vuːtən] 'foot' is attributed to an underlying distinction (/t/ vs. /d/), which is not directly observable (since learners have access only to surface forms), but must be inferred from its effect on surface forms. The grammar must be set up in such a way that /t/ and /d/ are neutralized in some contexts and distinct in others. Other instances of hidden structure that have been proposed in the literature include intermediate levels of representation in serial derivational frameworks, the assignment of segmental material into prosodic structure (feet; syllables; sub-syllabic constituents; weight-bearing units), and segmental feature specifications, including distinctions between full vs. underspecification and also language-particular assignments of phonological feature values to segments (Dresher 2004 Acquisition; Mielke 2004 Emergence, 2005 Modeling; Rice 2005 Liquid relationships). In all of these cases, the correct grammar cannot be found until the hidden structure has been established, while hypotheses about hidden structure cannot be evaluated until the corresponding grammar is constructed. In order to break into this circularity, the learner must have some independent means of establishing hypotheses about either the grammar or the hidden structure[10] (Tesar et al. 2003 Surgery; Apoussidou 2007 Learnability).

4.2.1 Proposals for learning underlying forms

Underlying forms are an especially challenging type of hidden structure to recover, since in principle there are infinitely many possible hypotheses about the underlying form of any given morpheme. Two assumptions have proven useful in helping the learner break into the system: (1) the prior stage of phonotactic learning provides an initial hypothesis about key aspects of the grammar, and (2) lexicon optimization, which favors underlying forms that are as close as possible to their corresponding surface forms, provides the learner with a set of initial hypotheses about underlying forms of a morpheme (Prince and Smolensky 1993/2004 Optimality Theory).

[10] A different approach to the learnability problem posed by hidden structure is to seek observable phonetic differences that would reveal, for example, syllabification of medial consonant clusters by duration cues (Maddieson 1985 Phonetic; Boucher 1988 Parameter; Tuller and Kelso 1991 Production) or weight bearing properties of rhyme consonants (Gordon 2004 Syllable weight). Furthermore, hidden structure is only a problem insofar as it actually influences phonological patterning, and in many instances the necessity of hidden structure for this purpose has been questioned or denied; see Prince 1983 'Relating' and Gordon 2002 'Factorial typology' on foot structure and stress placement; Steriade 1999 'Alternatives' on syllable boundaries and laryngeal contrast.

Proposals for establishing underlying forms typically rely on some form of the following strategy to establish initial hypotheses. First, if a morpheme never alternates in a particular feature, its underlying value is equal to its sole surface value (modulo robust interpretive parsing). This lets the learner establish a “skeletal frame” of invariant features for each morpheme (Inkelas 1995 Consequences; Tesar 2006 CogSci; Tesar and Prince 2007 Using phonotactics; see also Kenstowicz and Kisseberth 1977 Topics, chap 1 for discussion). Second, if a morpheme does alternate in a feature, we need some way of selecting an underlying value. This requires that the learner have available a set of hypotheses about possible underlying forms, and is able to evaluate which of these will lead to a grammar that is consistent with all of the known data.

One straightforward approach is to let the learner pick a value (arbitrarily) from among the set of attested surface values, and try to learn a grammar that goes along with this assumed UR (Kenstowicz and Kisseberth 1977 Topics, p. 33). If the first value that is chosen creates a ranking paradox so that it is not possible to learn a consistent ranking that covers all of the data, the learner retracts the hypothesis and tries a different value (Kager 1999 OT; Tesar et al. 2003 Surgery; Tesar and Prince 2007 CLS 39). In the case of Dutch voicing neutralization, the procedure works as follows: for invariant morphemes like [vut] ~ [vutən], the UR must be /vut/. For alternating morphemes like [bɛt] ~ [bɛdən], the learner has two choices: [–voice] and [+voice]. Suppose the learner starts by hypothesizing /bɛt/ ([–voice]). In order to validate this hypothesis, the learner seeks to construct a grammar that maps /bɛt+ən/ → [bɛdən], while at the same time mapping /vut+ən/ → [vutən]. These requirements are mutually incompatible, since the grammar must simultaneously allow intervocalic voicing (*VTV » Faith(voicing)) and maintain intervocalic voiceless stops (Faith(voicing) » *VTV). Thus, the hypothesis leads to inconsistency and can be rejected, leaving the learner to consider the hypothesis /bɛd/. In this case, it is no problem to learn a grammar that is compatible with the full set of known surface forms, since all that is required is that stops surface faithfully before sonorants and devoice elsewhere (Faith(voicing)/_[+sonorant] » *[+voice,–sonorant]). In fact, this is the ranking that the Dutch learner would already have from the prior stage of phonotactic learning, and a learner that makes maximal use of previous knowledge might favor this solution even if a consistent ranking could be learned with a different UR (Pater 2000 Nonuniformity; Prince and Tesar 2007 Using phonotactics). Thus, by trial and error the learner is able to arrive at a working combination of URs and grammar.

A related approach, proposed by Jarosz (2006 Rich lexicons), is to let the learner acquire lexical representations by entertaining all possible hypotheses of grammars and underlying forms simultaneously, using Maximum Likelihood Estimation to assign each combination a probability given the current set of data. In a case like Dutch, the learner considers URs with voiced and voiceless values, and grammars with final devoicing, intervocalic voicing, both processes, and neither process. As noted above, there is no combination of underlying forms that can generate the attested surface forms with intervocalic voicing or fully contrastive voicing, so the only grammar+UR combination that is assigned high probability after the model receives data from morphologically related forms is one that has an underlying voicing contrast and final devoicing.

These procedures work in the Dutch example, but they are not particularly efficient. Randomly trying out different feature values may require as many as 2^n guesses, where n = the number of alternating feature values in the lexicon, and as many runs of Recursive Constraint Demotion. Likewise, assessing probability distributions over all logically possible grammar+UR combinations is a computationally intensive task which is infeasible to carry out exhaustively in all but the simplest cases. Ultimately, the learner would benefit from a way of letting successful discovery of underlying values inform choices for other words. This can be seen most clearly in cases where multiple morphemes must have their underlying values set correctly before a consistent ranking can be found. Suppose the Dutch learner knew a number of plural forms with voicing ([bɛdən] 'beds', [hudən] 'hats', [hɑndən] 'hands', [krɑbən] 'crabs') at the time when the plural morpheme was learned, so that there are multiple alternating stems in the data. A consistent grammar cannot be found until every one of these morphemes is listed with an underlying [+voice] value. McCarthy (2005 Free ride) proposes a procedure by which decisions about underlying values may be extended to multiple morphemes at once, which could help guide the learner to this hypothesis.[11] In addition, the learner might make use of the fact that voicing contrasts are already known (from the prior stage of phonotactic learning) to surface faithfully only in pre-sonorant position to favor the value found in the plural. Finally, a more efficient learner might make use of the fact that some feature values are known never to contrast in any context on the surface, and are therefore unlikely to be useful in characterizing attested alternations (Dresher 2004 Acquisition).

Another weakness of these approaches is that progress is "all or nothing", since a hypothesized UR is deemed successful only once a consistent grammar is found that yields that attested surface form. This may be an overly stringent criterion in cases where morphemes participate in multiple alternations, since the learner may be able to make sense of certain aspects of the word but may not yet have sufficient data to arrive at a full analysis. For example, some Dutch nouns alternate not only in voicing, but also vowel length/quality: [bɑt] ~ [b**aː**dən] 'bath-SG./PL.', [smɪt] ~ [smedən] 'smith-SG./PL.'. It is plausible to think that learners may be able to establish the underlying voicing value even if they do not yet understand the (now lexically restricted) vowel alternation.[12] Apoussidou (2007, pp. 167–168 Learnability) proposes that knowledge of different underlying feature values of a morpheme are encoded separately, so that the learner need not arrive at a fully consistent ranking for all feature values simultaneously. The "all-or-nothing" criterion of success is also difficult to meet in cases where the choice of underlying values for one morpheme depends on the choice of values for another morpheme; in such cases, it is useful to allow the learner to focus on pairs of forms that differ by only a single morpheme at a time, in order to restrict the hypothesis space of possible modifications (Alderete et al. 2005 Contrast; Merchant and Tesar 2008 Learning).

A particularly challenging configuration concerns cases of three-way contrast: alternating morphemes A ~ B exist alongside both non-alternating A and non-alternating B. An example is provided by Turkish (Kaisse 1986 Locating; Inkelas 1995 Consequences):

[11] The proposals in Harrison and Kaun (2000 Pattern responsive) and McCarthy (2005 Free ride) are both intended to allow the learner to consider the possibility of extending alternations to morphemes that are not currently known to alternate. We suggest here that a similar strategy would be useful in handling morphemes that are known to alternate but have not yet been analyzed successfully.

[12] Support for this idea comes from the fact that Dutch vowel alternations are a relic of a formerly productive pattern of open syllable lengthening (Booij 1995, p. 88 Phonology of Dutch), which has become unproductive/lexically restricted in modern Dutch. The fact that many lexical items have maintained voicing alternations while losing vowel length alternations suggests that the underlying voicing value of individual morphemes was learned successfully, separately from the analysis of vowel length alternations.

(3)	sanat	~	sanat-ɯ	‘art-NOM./ACC.’
	kanat	~	kanad-ɯ	‘wing-NOM./ACC.’
	etyd	~	etyd-y	‘etude-NOM./ACC.’

Applying the reasoning above, the presence of non-alternating [t] and non-alternating [d] would straightforwardly lead the learner to posit URs such as /sanat/, /etyd/, which then requires a grammar that allows underlying voicing values to surface faithfully in all contexts (Faith(voice) » *VTV, *[+voice,–sonorant]/_[–sonorant]). The challenge is to infer underlying values for morphemes with alternating [t] ~ [d]: positing [–voice] would require a process of intervocalic voicing (incorrectly ruling out [sana**t**ɯ]), while positing [+voice] would require a process of final devoicing (incorrectly ruling out [ety**d**]). Numerous solutions to such configurations have been put forward in the literature, including underspecification of alternating segments to exempt them from faithfulness (Inkelas 1995 Consequences), or listing both values as underlying for alternating morphemes (Hooper 1976 Introduction; Kager, in press ‘Lexical irreg’). Such representations have proven effective in distinguishing many cases of alternating vs. non-alternating morphemes, but they come at a cost: the search space for underlying representations goes beyond the set of surface-observable feature values to include underspecified representations or even “overspecified” representations that include floating features or other structure that does not appear on the surface. (See Kenstowicz and Kisseberth 1977 ‘Topics’, chap. 1 for relevant discussion.)

One final challenge that is worth mentioning are cases in which the learner may wish to consider underlying values that are distinct from surface values, even in the absence of surface alternations. Kenstowicz and Kisseberth (1977 Topics) discuss an example from Yawelmani Yokuts in which the future suffix *-en*/*-on* undergoes rounding harmony to match preceding high vowels (*xil-en* ‘will tangle’ vs. *mut-on* ‘will swear’), but not non-high vowels (*bok’-en* ‘will find’), contrary to the usual pattern in the language of rounding harmony among vowels that agree in height. This fact suggests that the future suffix is underlyingly high, conditioning the expected rounding harmony with high vowels and then lowering. Unfortunately, the simplest version of this hypothesis—namely, that the suffix is underlyingly /-in/—is not tenable, since Yokuts has many short high vowels [i] and [u] that do not lower to [e], [o]. Kenstowicz and Kisseberth (following Kuroda 1967 Yawelmani) make use of the fact that Yokuts has long vowels, and that they are subject to two restrictions: they generally do not occur in closed syllables (*eːn, *oːn), and there are no long high vowels in suffixes of this type (*iː, *uː).[13] Putting these together, they posit that the future suffix *-en*/*-on* has an underlying long high vowel /iːn/, which undergoes rounding harmony with preceding high vowels, and then lowers to mid and shortens due to the coda consonant (see Kenstowicz and Kisseberth 1977 Topics, pp. 47–48 for details and arguments). This solution provides an elegant account of why the future suffix alternates in an unexpected way, but requires that learners consider underlying long vowels for morphemes that are always short on the surface. As Kenstowicz and Kisseberth point out, if such analyses are accepted, there are few (if any) criteria that can be imposed on possible

[13] Blevins (2004 Reconsideration) discusses several contexts in which long high vowels do occur in Yokuts and other Yawelmani dialects. The fact that long high vowels are possible in at least some contexts means that the relation between the harmony pattern and the surface phonotactics of Yawelmani is not as direct as it is sometimes portrayed in the literature, and calls into question (but does not preclude) the vowel-lowering analysis reviewed here.

divergences between underlying and surface forms. This makes it difficult to define formal procedures that can efficiently discover the full range of types of underlying forms that have been used in phonological analyses (see also Hockett 1955 Manual). A sensible heuristic (favored also by the principle of lexicon optimization) would be to favor underlying values that are as close as possible to attested surface values (Dresher 1981). Featural distance alone is not likely to be sufficient to guide the search, however, since the search space for underlying forms that differ by even a single feature value from the set of attested surface values may be quite large if floating features or abstract diacritic features are permitted.

Traditionally, considerations of learnability have not played a major role in helping to choose among possible theories of how to encode surface distinctions with underlying representations. We anticipate that as work proceeds on automated algorithmic discovery of underlying forms, the learnability ramifications of more complex representations may well be a more prominent factor in adopting one strategy over another.

4.2.2 Opacity

The example of the Yokuts future suffix discussed in the previous section is difficult not only because the hypothesized underlying long vowel never surfaces, but also because the interaction with rounding harmony is *opaque* (Kiparsky 1971 Historical, pp. 621–623): the suffix agrees in rounding with a preceding high vowel, but surfaces as a [–high] vowel which would otherwise be exempt from [+high] rounding harmony: /t’ujt’uj-iːn/ → [t’ujt’ujon] ‘will shoot repeatedly’ (cf. [hud-al]/*[hud-ol] ‘might recognize’). This is an example of counterbleeding opacity: harmony occurs even though an independent process intervenes, removing the apparent motivation for the change. At the same time, the suffix fails to agree in rounding with preceding round vowels that do match in height: /bok’-iːn/ → [bok’en]/*[bok’on] ‘will find’. This is an example of counterfeeding opacity: lowering of /iː/ to [e] creates a mid vowel that would ordinarily be subject to harmony, but it fails to harmonize due to its underlying [+high] status.

(4) Opacity in the Yokuts future suffix

UR	/t’uyt’uy-iːn/	/bok’-iːn/
Rounding harmony	t’uyt’uyuːn	*n.a.*
Lowering of high vowels	t’uyt’uyoːn	bok’eːn
Closed syllable shortening	[t’uyt’uyon]	[bok’en]

In both cases, the relation between phonotactics and alternations is disrupted. In the case of counterbleeding opacity, the learner encounters apparently unmotivated alternations that prior knowledge of surface phonotactics cannot help to explain (the context for the alternation is not surface-apparent), while in the case of counterfeeding opacity, the learner encounters surface exceptions that stand in the way of learning the alternation in the first place (i.e. it is not surface-true). The intuition has often been expressed in the literature that these features of opacity must be an obstacle to learning opaque interactions (Kenstowicz and Kisseberth 1977, p. 169; Hock 1991 Principles, chap. 11).

In many cases, it is plausible to suppose that the learner is aided by a large number of forms in which just one of the two processes applies, allowing a certain amount of grammatical learning based on unambiguous data (Bermúdez-Otero 2003 Acquisition). For example, in Yokuts, the non-future suffix *-hin*/*-hun*, the perfective suffix *-mi*/*-mu*, the future passive suffix *-nit*/*-nut* and the dubitative suffix *-al*/*-ol* all show the general pattern of height-conditioned rounding harmony. This could conceivably strengthen the conviction of the learner that the observed alternations are in fact all motivated by the same phonological constraints, and help guide the learner to posit abstract levels of representation in which the same conditions are present for the opaque cases.

Even when simpler unambiguous cases are available, however, the task of learning opaque interactions between multiple processes is necessarily more difficult than the task of learning a single alternation. Indeed, closer scrutiny into what speakers actually extract from data involving opaque interactions may shed light on the workings of the system (Mielke, Armstrong and Hume 2003 Looking). For example, Sanders (2003 Opacity) tested the willingness of Polish speakers to generalize an opaque vowel-raising process to novel words, and found that the alternation, though amply attested in the lexicon, was not extended productively. On the other hand, Poliquin (2006 Canadian French, pp. 136–143) found that Canadian French speakers readily apply an opaque vowel harmony process to low frequency and novel words. Clearly further experimental work on the synchronic productivity of opaque processes will be an important source of evidence concerning whether (and how) speakers learn them.

Another important traditional source of evidence about what is learned comes from language change. It has long been observed that opaque interactions are unstable, and are frequently reanalyzed such that both processes apply transparently, or one of the processes is lost (Kiparsky 1965 Phonological Change; King 1969 Historical, pp. 87–101). Hansson and Sprouse (1999 Factors) contrast the fate of rounding harmony in a later generation of Yokuts speakers, observing that harmony among high vowels is preserved (/ʔukn-hin/ → [ʔukun-h**u**n] 'drink-NON-FUT.') while harmony among non-high vowels is lost (/woːn-k'a/ → [won-k'**a**]). The difference appears to be in how the two processes interacted with vowel lowering. As noted above, lowering counterbleeds high vowel harmony causing it to apply in more cases than expected, while it counterfeeds low vowel harmony and creates surface exceptions. Hansson hypothesizes (consistent with claims by Kiparsky, King, and others) that harmony among high vowels was easier to learn in the original system because it applied consistently in at least a subset of the relevant contexts. This provides another piece of evidence that bootstrapping from a subset of the data that shows transparent and reliable application may provide an important entry into the system.

One additional factor that appears to facilitate the learning (and creation) of opaque rule orderings is the fact that counterbleeding interactions frequently reduce alternations, leading to greater paradigm uniformity (Kiparsky 1972 Explanation; King 1973 Rule insertion; Kenstowicz and Kisseberth 1977 Topics, pp. 163–164; Burzio 1996 Surface; Kenstowicz 1997 Base identity). McCarthy (1998 Occultation) argues for independent reasons that learners must be a biased to place paradigm uniformity (output-output faithfulness) constraints at the top of the ranking, above markedness constraints. This correctly predicts that learners should easily be able to analyze—or may even accidentally create—opaque interactions that eliminate paradigmatic alternations, as in the Greek example described in section 4.1.

It should also be emphasized that the fact that speakers frequently stop applying opaque processes should not be taken as evidence that learners fail to notice them entirely. In fact, there is reason to think that when learners are confronted with conflicting data caused by counterbleeding interactions, they seek to explain the competition by exploring complex and detailed conditioning environments; we return to this issue in section 4.5.

4.3 Theories for learning alternations II: approaches using surface mappings

An alternative approach to encoding alternations within paradigms is as relations among surface forms. Returning to the Dutch example [bɛt] ~ [bɛdən] 'bed-SG./PL.', one could observe that stem-final [d] in the plural corresponds with [t] in the singular (though not always the reverse), and encode this directly as a relation between surface forms. One common approach to limiting phonological processes to relations between surface forms is to require that the underlying form match one attested surface allomorph (Harris 1942 Morpheme alternants, 1951 Methods, p. 308 fn. 14; McCawley 1967 Sapir; Vennemann 1974 Concreteness; Hooper 1976 Introduction; Kenstowicz and Kisseberth 1977 Topics, pp. 28–33). In other theories, alternations are simply built in to the statement of the morphological mapping: [Xd-ən] in the plural → [Xt] in the singular (Zwicky 1985 Inflection; Wurzel 1987 Paradigmenstrukturbedingungen; Bochner 1993 Simplicity; Barr 1994 Lexical; Albright and Hayes 2002 Modeling). Alternatively, work within the framework of Optimality Theory has proposed to capture such surface relations using the machinery of output-output correspondence constraints (Burzio 1996 surface; Russell 1999 MOT; Cole and Hualde 1998 Lexical acquisition; MacBride 2004 Constraint-based).

When underlying forms (or inputs to morphological mappings) are limited to surface forms, the search space for underlying forms is greatly reduced. This is not guaranteed to reduce the learning challenge, however, since the learner must instead find reliable implicational relations between surface forms. A learner of Dutch, for example, would need to learn that a voiceless obstruent in the plural ([vu**t**-ən] 'foot-PL.') reliably corresponds to a voiceless obstruent in the singular ([vu**t**] 'foot.SG.'), but the reverse does not hold ([bɛ**t**] ~ [bɛ**d**-ən] 'bed.SG./PL.', not *[bɛ**t**-ən]). In learning the predictors of voicing, two kinds of search are useful: a search for phonological contexts that frequently accompany voicing, and a search for those surface forms that most reliably reveal voicing.

First, learners may search for phonological contexts that are correlated with the difference in voicing between [vu**t**-ən] and [bɛ**d**-ən]. Ernestus and Baayen (2003 Unpredictable) show that voicing of stem-final obstruents can be predicted to a significant extent based on the place and manner of the segment in question, as well as features such as the preceding vowel length. They provide experimental evidence that speakers are able to use these lexical trends to predict the probability of alternations in nonce words. (We return to the issue of lexical gradience below in section 4.5). One procedure for discovering reliable predictors of an alternation is what Albright and Hayes (2002 Modeling, 2003 Rules), building on a proposal sketched by Pinker and Prince (1988 Language and connectionism), call the **minimal generalization** approach: the learner compares pairs of morphologically related surface forms to determine what they have in common and what varies between the two forms. For example, a Dutch learner confronted with the pair [vut] ~ [vutən] 'foot-SG./PL.' would align the material in the two forms to discover that they differ only in the addition of a suffix (∅ → ən/__#), while alternating form like [bɛt] ~

[bɛdən] 'bed-SG./PL.' differ both in voicing and the addition of a suffix (t → dən/__#). By comparing additional pairs such as [rat] ~ [radən] 'wheel-SG./PL.', the learner attempts to extract phonological features that statistically favor alternation or non-alternation. Based on just these three items ([vut] vs. [bɛt], [rat]) , the height and rounding of the preceding vowel look like they might be reliable indicators, with voicing alternations occurring after [–high] or [–round] vowels. Consideration of more data would reveal that this particular correlation turns out not to be particularly strong in the Dutch lexicon, but other features such as vowel length are strongly correlated with voicing alternations (Ernestus and Baayen 2003 Predicting; Kerkhoff 2007 Acquisition, pp. 96–104). Other algorithmic approaches to identifying predictive contexts include decision tree-based approaches (Breiman et al. 1984 CART; Ling and Marinov 1993 Answering; Gildea and Jurafsky 1996 Learning bias), the Analogical Modeling of Language (AML: Skousen 1989 Analogical; Eddington 2003 Issues), TiMBL (Daelemans 2000 TiMBL), and the Generalized Context Model (Nosofsky 1986 Attention, 1990 Relations; Nakisa, Plunkett and Hahn 1997 Cross-linguistic).

A second type of information that can help ensure accurate inferences based on surface forms is the knowledge that some forms are better than others at revealing surface contrasts. For example, in Dutch nouns it is clear that the plural is a better source of information than the singular about the voicing of stem-final obstruents, since the singular undergoes final devoicing while the plural maintains voicing contrasts. Thus, a learner might wish to learn about asymmetries in the predictive power of different members of the paradigm. Albright (2002 Identification) proposes a procedure in which learners compare the reliability of mappings based on different available surface forms by using the minimal generalization algorithm to construct grammars using each part of the paradigm as an input, and evaluating the accuracy of the resulting grammars. In this way, the learner can discover that some parts of the paradigm undergo more neutralizations than others, and can subsequently focus on just those mappings that are known to have high predictive value. As has long been noted (e.g., Kenstowicz and Kissberth 1977 Topics, pp. 28–33), theories that operate on surface allomorphs are much more restrictive than those that operate on more abstract underlying representations. A potential advantage of this restrictiveness is that it greatly simplifies the learning task, since the learner need only identify those parts of the paradigm that tend to be most informative in the language rather than comparing all forms of all words to locate contrastive values on a morpheme-by-morpheme basis.

There is reason to believe that learners do indeed focus on particular parts of the paradigm that are characteristically most informative. A particularly revealing source of evidence comes from cases of "consistent inheritance", in which idiosyncratic properties of one paradigm member are carried over to other paradigm members. Spanish provides a telling example. Many Spanish verbs show alternations between a velar stop in some forms and ∅ in others:

(5) Spanish velar alternations

salir 'to leave'	Present indicative	Present subjunctive
1SG	sal**g**-o	sal**g**-a
3SG	sal-e	sal**g**-a
1PL	sal-imos	sal**g**-amos

An approach using underlying forms might posit an underlying /g/ that deletes before front vowels: /salg-e/ → [sale], producing paradigms in which [g] is retained only before back vowels (Harris 1969 Spanish phonology). An approach based on surface mappings would instead rely on implicational relations among surface forms: the present subjunctive matches the form found in the 1sg present indicative.. Although this statement misses the relation between presence of [g] and the following vowel quality, it makes a much more general prediction: the subjunctive should always resemble the 1SG indicative. This prediction is in fact correct: in verbs where the 1SG indicative differs from the remaining indicative forms in other idiosyncratic ways, the present subjunctive consistently inherits the properties of the 1SG indicative (Maiden 2005 Morphological autonomy).

(6) Idiosyncratic alternations in Spanish

caber 'to fit'	Present indicative		Present subjunctive	
1SG	quep-o	[**kep**o]	quep-a	[**kep**a]
3SG	cab-e	[kabe]	quep-a	[**kep**a]
1PL	cab-emos	[kabemos]	quep-amos	[**kep**amos]

The phenomenon of consistent inheritance is sometimes referred to as parasitic or Priscianic derivation (Matthews 1972 Inflectional; Aronoff 1994 Morphology). That these resemblances are not accidental is shown by the fact that speakers appear to actively enforce them, analogically replacing exceptional forms with novel ones that conform to the inherited relationship. This may be taken to indicate that speakers learn systematic relations among particular surface forms within the paradigm (Zwicky 1985 Inflection; Stump 2001 Inflectional).

4.4 The subset problem in alternations: optional rules

As with phonotactics, learning alternations from positive evidence alone may pose a subset challenge. A particularly interesting case of this, pointed out by Dell (1981 Learnability), involves the problem of learning whether a rule is optional or obligatory. Dell observes that final obstruent + liquid clusters may optionally be simplified in French: /bukl/ 'buckle' optionally pronounced [buk]. This simplification is not possible for words ending in obstruent + nasal or obstruent + clusters: /ritm/ → [ritm], *[rit] 'rhythm', /fiks/ → [fiks], *[fik] 'fixed'. The challenge for a learner restricted to positive evidence is to determine, based on positive examples like [ritm], that [rit] would not be a grammatical variant. This is fully parallel to the example discussed above in which the learner, presented solely with positive examples of [pa], [ba], and [ma] must infer that [m̥a] is not grammatical. Dell proposes that learners employ an explicit heuristic principle of adopting the most restrictive grammar possible. As discussed in section 3.3, one way to implement a restrictiveness bias in Optimality Theory is to favor rankings of Markedness constraints over Faithfulness constraints. It is important to note that in this case, however, the challenge is to demand greater faithfulness in the absence of explicit evidence of alternations (i.e., the learner must assume that nasals and obstruents may not be deleted, but must be pronounced faithfully). It appears that the most general solution to the subset problem is one that employs a principle such as Entropy (Riggle 2006 Entropy) or Maximum Likelihood Estimation (Jarosz, in press, Restriveness), which rely on metrics that bear an invariant relation to restrictiveness, rather than an approach that attempts to regulate Markedness and Faithfulness rankings directly.

4.5 The gradience problem in alternations

As with static phonotactics, alternations do not apply exceptionlessly. In many cases, the alternation is lexically restricted: some morphemes consistently undergo them, while others are consistently immune. The Turkish example discussed above could be seen as a case of this: final devoicing and intervocalic voicing are enforced for morphemes [kanat] ~ [kanad-ɯ] 'wing-NOM./ACC.', but not for morphemes like /sanat/ 'art' or /etyd/ 'etude'. Numerous studies have used wug tests to explore speakers' knowledge of lexically gradient alternations. In general, it appears that when a process has exceptions and applies with different probability to words of different phonological shapes, speakers' behavior on wug words tracks these differences (Zuraw 2000 Patterned; Albright, Andrade and Hayes 2001 Segmental; Pierrehumbert 2006 Unnatural; Hayes and Londe 2006 Stochastic; and many others).

As noted above, lexically gradient processes pose a learning challenge because they create inconsistencies that are difficult to capture with a single constraint ranking. Under a theory that attempts to augment underlying representations to reconcile all morphemes with a single grammar, a standard solution is to use diacritics to mark certain morphemes as exceptions to particular rules/constraints. Such a theory attributes no particular significance to the fact that a particular rule has exceptions in some morphemes; the existence of exceptions is simply a static fact about the lexicon, and no explicit mechanism is provided for speakers to extend gradience to novel items in a wug test. One plausible assumption would be that when speakers are given incomplete information about a novel morpheme in the context of a wug test, they examine the lexicon to assess the probability of different underlying representations (Schütze 2005 Thinking); a procedure along these lines is proposed by Harrison and Kaun (2000 Pattern-responsive).

An alternative approach that recognizes and reifies the side-by-side existence of different patterns is to abandon the goal of finding a single consistent grammar, instead allowing morphemes to be associated with different constraint rankings (Ito and Mester 2002 Phonological lexicon; Anttila 2002 Morphologically conditioned; Inkelas and Zoll 2003 Grammar dependence; Pater 2000 Nonuniformity; Becker 2008 Phonological Trends). For instance, Pater (2008a Morpheme-specific) proposes that when learners are confronted with inconsistent pairs such as *kana**t** ~ kana**d**-ɯ* vs. *ety**d** ~ ety**d**-y*, they seek to resolve the conflict by finding a constraint that may be ranked differently for different morphemes. In this case, faithfulness for voicing could be ranked high for words like /etyd/, and ranked low for words like /kanad/. Such proposals make use of the fact that the search space of rankings, while large, is easier to define and search than the space of possible lexicons employing underspecified and augmented underlying representations. In addition, analyses in terms of competing rankings provide a mechanism for encoding the fact that different words behave differently directly in the grammar, providing a natural mechanism for gradient generalization to novel items (Pater 2008a Morpheme-specific) For instance, Becker (2008 Phonological) proposes that when learners discover that a constraint is variably ranked, they keep track of the number of morphemes that obey each ranking and can use this knowledge to estimate the probability with which a novel morpheme should obey a particular ranking.

Yet another approach to lexically gradient alternations is to use the grammar to encode knowledge of the probability of participating in the alternation, and the lexicon to encode the behavior of individual lexical items. Zuraw (2000 Patterned) uses the Gradual Learning Algorithm to allow conflicting data from lexically gradient processes to lead to non-categorical rankings, which may then be generalized to novel items at ratios matching the rate of alternation in the training data. In order to capture the fact that existing (known) morphemes are generally consistent in their behavior, it is proposed that speakers rely on these memorized word-specific knowledge which blocks the variability that the grammar would otherwise produce. A similar approach can be seen in the minimal generalization model of Albright and Hayes (2003 Rules), which encodes competing lexically gradient patterns by means of probabilistic rules, and relies on word-specific knowledge to ensure that known words are inflected consistently.

5. What doesn't have to be learned? The issue of UG

An important recent development in the study of gradient processes is the possibility that not all statistical trends are equally learnable. For example, Becker, Ketrez and Nevins show that the probability of voicing alternations in Turkish is correlated with a number of features in the surrounding context, such as the place of articulation, the length of the word, and the preceding vowel quality. However, they argue that in this case, wug test data does not mirror the lexical trends as closely as in the examples cited above: Turkish speakers are sensitive to the role of consonant place and word length, but do not appear to take vowel features into account when deciding on the probability of voicing. This highlights the fact that progress is modeling gradient processes is likely to require not only better statistical models of learning from the lexicon, but also a better understanding of which trends speakers choose to encode, and at what level of granularity.

Some of the learning models mentioned above are sharply inductivist, attempting to find the right phonological grammar using very little a priori knowledge, perhaps limited to just a feature inventory and the learning principles themselves. We think the development of such systems is a good research strategy—not because the ultimate right answer to the problem of phonological learning is necessarily a purely inductivist one, but because inductivist approaches can be used to gain insight into UG proposals.

A pioneering contribution in this area is Gildea and Jurafsky (1996 Learning bias), which sought to develop a formal system that, given input/output pairs, could learn appropriate phonological rules to relate the two. They adopted as their baseline algorithm "OSTIA", a procedure for discovering finite state transducers invented by Oncina, García, and Vidal (1993 Learning). Applying OSTIA to English phonological data, Gildea and Jurafsky found that the algorithm could learn versions of rules like Flapping only after they had augmented it with three further principles, which are at least tacitly present in almost any phonological theory: "**Faithfulness** (underlying segments tend to be realized similarly on the surface), **Community** (Similar segments behave similarly), and **Context** (Phonological rules need access to variables in their context)" (p. 497). One potential interpretation of this is that the three abstract principles must *necessarily* be part of phonological theory, since learning would be impossible without them. Of course, Gildea and Jurafsky are cautious on this point, since it is possible that some other primitive inductive system might solve the problem as well, or that the three principles might themselves be learnable.

A similar research strategy is adopted in the phonotactic learner of Hayes and Wilson (2008 Maximum Entropy), mentioned above. Hayes and Wilson find that their basic inductive system is defeated by nonlocal phonological phenomena such as stress and harmony, which can involve segments that are at some distance from one another. They find that such systems can be learned when the phonological theory assumed is augmented to include standard generative phonological formalisms for such phenomena, specifically metrical grids and autosegmental tiers. The crucial difference these representations make is that they provide local formal characterizations of surface-nonlocal configurations, permitting phonotactic patterns to be learned that would otherwise be inaccessible, being expressible only as hypotheses that occupy huge, unsearchable hypothesis spaces.[14] It is plausible to imagine that a number of other elements of phonological theory would likewise facilitate learning—and also possible that some proposals actually hinder it, by expanding the hypothesis space with no compensating gain in access to the useful hypotheses.

Learnability studies complement experimental work that seeks to find direct evidence for UG principles. One such type of experiment assess whether speakers have phonotactic preferences that distinguish forms that are equally unattested in their language: for instance, the form [lbɪf] (monosyllabic) is illegal in English, but has a more severe violation of sonority sequencing principles (e.g. Sievers 1901 Grundzüge) than [bdɪf]. Berent et al. (2007, 2008; see also Pertz and Bever 1975 Sensitivity) find that in various tasks, English speakers act in ways indicating that [lbɪf] is less well-formed [bdɪf], and cautiously suggest that this reflects sonority sequencing as an *a priori* principle that influences phonetic judgments, independently of whatever phonological principles are learned from exposure to data.

This result can be evaluated further if we use computational learning models. The idea is that perhaps, contrary to initial assumptions, the [lb] - [bd] *is* learnable, being implicit in the ample overt evidence that English onset clusters do respect sonority sequencing in some general sense. Albright (2007 ms., Natural classes) conducts further experiments on initial sonority sequencing, modeling his results with both an analogical model similar to that of Bailey and Hahn (2001 Wordlikeness) and his own phonotactic learning algorithm; neither model predicts all the sonority based acceptability differences in the experimental data, thus tentatively supporting the conclusion that sonority sequencing embodies a priori knowledge.[15] Although this specific conclusion may be overturned by subsequent advances in automated learning, it illustrates a more general principle: computationally implemented learning models provide a

[14] Plainly, this is a tentative result, since it depends on the claim that no other learning mechanism would be able to find the crucial generalizations without the a priori provision of tiers and grids; see Goldsmith and Xanthos, 2006.

[15] As far as we can tell, the phonotactic learner proposed by Hayes and Wilson (2008 Maximum Entropy) likewise cannot project correct native speaker judgments about sonority sequencing simply by generalizing from the attested English data.

concrete estimate of what we can responsibly assume that learners may extract from the data, and guide the researcher towards aspects of phonological patterning that appear to be difficult to extract from the data.

Another important strategy for obtaining grammaticality intuitions that could not have come from the acquisition data is to construct pairs of entirely new miniature languages that differ in crucial respects and compare people's ability to learn them. The contrasting properties of the language pairs must be uncued (i.e., statistically neutral) in the native language of the experimental subjects. Wilson (2006 Learning) set up such an experiment, based on the well-known typological observation that palatalization of velars is favored in the environment before high front vowels relative to before lower ones (Chen 1972 Formal expression). In his experiment, subjects showed some tendency to generalize the rule k → tʃ / ___ e also to cover k → tʃ / ___ i, but not in the opposite direction, a pattern for which their prior (monolingual) experience with English provides no direct evidence. This can be taken, at a very simple level, as a "UG in action" result, but Wilson pursues the issue more intensively by asking what sort of UG, and what kind of learning model might project the result from deeper principles. Wilson's view is that the (perhaps innate) principle at stake is Paradigm Uniformity, taken at the phonetic level (Steriade 2000 Phonetics): speakers are a priori more willing to tolerate alternation between phonetically similar pairs than phonetically distant one. In the present case, [ke] is further from [tʃe] than [ki] is from [tʃi] (due to the greater burst noise in [ki] than [ke]; Guion 1998 Role of perception), so speakers are a priori more willing to tolerate [ki] ~ [tʃi] alternation than [ke] ~ [tʃe] alternation. The most striking aspect of Wilson's study is the final step, which is to construct an implemented model of what the experimental subjects were doing when they learned the constructed languages. The model is partly inductive, and partly the application of the innate principle of phonetic Paradigm Uniformity. Formally, Wilson implements this as a constraint based Maximum Entropy model, in which the weighting of the constraints depends on a prior term that governs the degree to which the constraint weight responds to data during learning. The weights express the final learned grammar, whereas the prior terms express the effect of UG principles on how grammars are learned.

Wilson's work formalizes the idea that UG principles may not always be absolute, but rather can express **learning biases**, whereby the principles guide but do not absolutely dictate the form of the grammar that is shaped on exposure to data. While the bias Wilson examines is phonetic, the variety of cases to be explored and modeled is far wider. Thus, Moreton (2008 Analytic bias) offers experimental data suggesting that speakers more easily pick up phonological patterns expressible in terms of identity (here, of vowel quality, or perhaps of height) than other patterns of equal phonetic naturalness (in Moreton's study, a vowel height/consonant voicing correlation).

6. Conclusion

In this chapter, we have attempted to highlight some of the challenges that learners face in analyzing phonological distributions and alternations. In many cases, our current state of knowledge is clearly still quite preliminary, based on schematic and idealized examples . Nonetheless, we believe that significant progress that has been made in formalizing the problem and providing concrete frameworks for solving it since 1955, when Hockett declared that "[w]e know of no set of procedures by which a Martian, or a machine, could analyze a phonologic system" (Manual, p. 147). Furthermore, we anticipate that as computational resources and power expands, current proposals may be subjected to broader and more realistic testing and use of implemented learners will become more widespread, allowing considerations of learnability to play a more central role in guiding phonological theory.

References

Aksu-Koç, A. and D. I. Slobin (1985)). Acquisition of Turkish. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 1. The data*, pp. 839–878. Hillsdale, NJ: Lawrence Erlbaum Associates.

Albright, Adam, and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the 2002 Workshop on Morphological Learning, Association of Computational Linguistics*, ed. Michael Maxwell, 58-69. Philadelphia: 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.

Albright, Adam. 2002. The identification of bases in morphological paradigms. Doctoral dissertation, University of California, Los Angeles.

Albright, Adam (in press) Gradient phonological acceptability as a grammatical effect. *Phonology*.

Albright, Adam (2007) Natural classes are not enough: Biased generalization in novel onset clusters. MIT ms.

Albright, A., A. E. Andrade, and B. Hayes (2001). Segmental environments of Spanish diphthongization. In A. Albright and T. Cho (Eds.), *UCLA Working Papers in Linguistics, Number 7: Papers in Phonology 5*, pp. 117–151.

Albro, D. M. (1998). *Evaluation, implementation, and extension of Primitive Optimality Theory*. Master's thesis, UCLA. http://www.linguistics.ucla.edu/people/grads/albro/ma.pdf

Albro, D. M. (2005). *Studies in Computational Optimality Theory, with Special Reference to the Phonological System of Malagasy*. Ph. D. thesis, UCLA. http://www.linguistics.ucla.edu/people/grads/albro/diss.pdf

Alderete, J., A. Brasoveanu, N. Merchant, A. Prince, and B. Tesar (2005). Contrast analysis aids the learning of phonological underlying forms. In J. Alderete, C. hye Han, and A. Kochetov (Eds.), *Proceedings of the 24th West Coast Conference on Formal Linguistics*, pp. 34–42. Somerville, MA: Cascadilla Proceedings Project.

Anttila, Arto. 1997a. Variation in Finnish phonology and morphology. Doctoral dissertation, Stanford University, Stanford, Calif.

Anttila, Arto. 1997b. Deriving variation from grammar: A study of Finnish genitives. In *Variation, change and phonological theory*, ed. Frans Hinskens, Roeland van Hout, and Leo Wetzels. Amsterdam: John Benjamins. Rutgers Optimality Archive ROA-63, http://ruccs.rutgers.edu/roa.html.

Anttila, A. (2002). Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory* 20(1), 1–42. ROA 425

Apoussidou, Diana. (2007). *The Learnability of Metrical Phonology*. Utrecht: LOT.

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes*. MIT Press.

Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language* 44:568–591.

Baker, W. and B. L. Derwing (1982). Response coincidence analysis as evidence for language acquisition strategies. *Applied Psycholinguistics 3*, 193–221.

Baković, Eric. (2007). A revised typology of opaque generalisations. *Phonology* 24, 217–259.

Baroni, M. (2000). *Distributional Cues in Morpheme Discovery: A Computational Model and Empirical Evidence*. Ph. D. thesis, University of California, Los Angeles.

Barr, Robin. (1994). *A lexical model of morphological change*. Ph. D. thesis, Harvard.

Becker, Michael (2008) Phonological trends in the lexicon. UMass dissertation.

Beddor, P. S., J. D. Harnsberger, and S. Lindemann (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics* 30, 591–627.

Berent, Iris, Donca Steriade, Tracy Lennertz, and Vered Vaknin. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104:591–630.

Berent, Iris, Lennertz, Tracy, Jun, Jongho, Moreno, M., A., & Smolensky, Paul. 2008). Language universals in human brains. Proceedings of the National Academy of Sciences, 105, 5321-5325

Berko, J. (1958). The child's acquisition of English morphology. *Word* 14, 150–177.

Bermúdez-Otero, R. (2003). The acquisition of phonological opacity. In J. Spenader, A. Eriksson, and O. Dahl (Eds.), *Variation within Optimality Theory: Proceedings of the Stockholm Workshop on 'Variation within Optimality Theory'*, pp. 25–36. Stockholm: Department of Linguistics, Stockholm University.

Berwick, R. C. (1986). Learning from positive-only examples: The subset principle and three case studies. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach: Volume II*, pp. 625–645. Los Altos, CA: Kaufmann.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Blevins, Juliette and Andrew Garrett (2004). The evolution of metathesis. In B. Hayes, R. Kirchner, and D. Steriade (Eds.), *Phonetically based phonology*. Cambridge University Press.

Bloomfield, L. (1933). Language. Chicago: University of Chicago Press.

Bochner, H. (1993). *Simplicity in Generative Morphology*. Berlin: Mouton de Gruyter.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.

Boersma, Paul, and Joe Pater (2008 ms.) Convergence properties of a gradual learning algorithm for Harmonic Grammar. ROA 970.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. In *Institute of Phonetic Sciences, University of Amsterdam, Proceedings* 21, 43–58.

Boersma, Paul. 1998. Functional Phonology: Formalizing the interactions between articulatory and perceptual drives. Doctoral dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.

Boersma, P. and C. Levelt (2000). Gradual constraint-ranking learning algorithm predicts acquisition order. In *Proceedings of Child Language Research Forum*, Volume 30, Stanford, pp. 229–237. CSLI Publications.

Booij, G. (1995). The Phonology of Dutch. Oxford: Clarendon Press.

Boucher, V. (1988). A parameter of syllabification for vstopv and relative-timing invariance. *Journal of Phonetics 16*, 299–326.

Brent, M. and T. Cartwright (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61, 93–125.

Burzio, L. (1996). Surface constraints versus Underlying Representation. In J. Durand and B. Laks (Eds.), *Current trends in phonology: Models and methods*, pp. 97–122. CNRS, Paris, and University of Salford: University of Salford Publications.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Chen, M. (1972). On the formal expression of natural rules in phonology. *Journal of Linguistics 9*, 209–383.

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

Clahsen, H., F. Aveledo, and I. Roca. 2002. The development of regular and irregular verb inflection in Spanish child language. *Journal of Child Language* 29:591–622.

Clahsen, H., Prüfert, P., Eisenbeiß, S., and Cholin, J. (2002). Strong stems in the German mental lexicon: Evidence from child language acquisition and adult processing. In Kaufmann, I. and B. Stiebels (eds), *More than Words. Festschrift for Dieter Wunderlich*, 91–112. Akadamie Verlag, Berlin.

Coetzee, A. and J. Pater (2008). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory 26*, 289–337.

Cole, J. and J. Hualde (1998). The object of lexical acquisition: A UR-free model. In *CLS 34. Chicago Linguistic Society* 34. *The Panels*: 447-458

Coleman, John, and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Computational Phonology, Third Meeting of the ACL Special Interest Group in Computational Phonology*, 49–56. Somerset, N.J.: Association for Computational Linguistics.

Curtin, Suzanne and Kie Zuraw (2002). Explaining Constraint Demotion in a Developing System. In Anna H.-J. Do, Laura Domínguez, and Aimee Johansen, editors, BUCLD 26: Proceedings of the 26th annual Boston University Conference on Language Development. Cascadilla Press.

Daelemans, W., J. Zavrel, K. Van der Sloot, and A. Van den Bosch (2000). *TiMBL: Tilburg memory based learner reference guide 3.0.* Report 00-01, Computational Linguistics Tilburg University.

de Marcken, Carl (1996). *Unsupervised Language Acquisition.* Ph. D. thesis, MIT.

Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12:31–37.

Derwing, B. L. and W.J. Baker (1986). Assessing morphological development. In P. Fletcher and M. Garman (Eds.), *Language acquisition: Studies in first language development* (2nd ed.)., pp. 326–338. Cambridge University Press.

Dresher, B. E. (1981). On the learnability of abstract phonology. In C. L. Baker and J. J. McCarthy (Eds.), *The Logical Problem of Language Acquisition*, pp. 188–210. MIT Press.

Dresher, B. E. (2004). On the acquisition of phonological representations. In Proceedings of the First Workshop on Psycho-Computational Models of Human Language Acquisition (Held in cooperation with COLING-2004, Geneva, 28 August 2004), pp. 41–48.

Dresher, B. Elan, and Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* 20:421–451.

Eddington, David (2003) Issues in modeling language processing analogically. *Lingua* 114, pp. 849—871.

Eisner, Jason. 1997. Efficient generation in primitive Optimality Theory. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 313–320. East Stroudsburg, Penn.: Association for Computational Linguistics.

Eisner, Jason. 2001. Expectational semirings: Flexible EM for finitestate transducers. In *Proceedings of the ESSLLI Workshop on Finite-State Methods in NLP (FSMNLP)*, ed. G. van Noord.

Eisner, Jason. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 1–8. East Stroudsburg, Penn.: Association for Computational Linguistics.

Ellison, T. Mark. 1994. Phonological derivation in Optimality Theory. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, 1007–1013.

Ernestus, M. and R. H. Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79, 5–38.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research 46*, 127–139.

Frank, Robert, and Shyam Kapur. 1996. On the use of triggers in parameter setting. *Linguistic Inquiry* 27:623–660.

Friederici, Angela D. & Jeanine E. Wessels (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics* **54**: 287-295.

Frisch, Stefan A., and Bushra A. Zawaydeh. 2001. The psychological reality of OCP-place in Arabic. *Language* 77:91–106.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22:179–228.

Frisch, Stefan A., Nathan R. Large, and David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42:481–496.

Gibson, Edward, and Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25:407–454.

Gildea, Daniel, and Daniel Jurafsky. 1996. Learning bias and phonological rule induction. *Computational Linguistics* 22:497–530.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics 27*, 153–198.

Goldsmith, John, and Aris Xanthos. 2006. Learning phonological categories. Ms., Department of Linguistics, University of Chicago. Downloaded May 21, 2007 from http://hum.uchicago.edu/~jagoldsm/Papers/phonolcat.pdf.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120.

Goldwater, Sharon. 2007. Nonparametric Bayesian models of lexical acquisition. Doctoral dissertation, Brown University.

Gordon, Matthew. 2002. A factorial typology of quantity insensitive stress. *Natural Language and Linguistic Theory* 20, 491-552

Gordon, Matthew. 2004. Syllable weight. In *Phonetically based phonology*, ed. Bruce Hayes, Robert Kirchner, and Donca Steriade, 277–312. Cambridge: Cambridge University Press.

Greenberg, Joseph H., and J. J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English. *Word* 20:157–177.

Grünwald, Peter, In Jae Myung, and Mark Pitt, ed. 2005. *Advances in minimum description length: theory and applications*. Cambridge, Mass.: MIT Press.

Guion, S. G. (1998). The role of perception in the sound change of velar palatalization. *Phonetica 55*, 18–52.

Hale, M. and C. Reiss (1998). Formal and empirical arguments concerning phonological acquisition. *Linguistic Inquiry 29*, 656–683.

Hale, M. and C. Reiss (2003). The subset principle in phonology : why the tabula can't be rasa. *Journal of Linguistics 39*, 219–244.

Halle, M. (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language. In M. Halle, J. Bresnan, and G. Miller (Eds.), *Linguistic Theory and Psychological Reality*, pp. 294–303. MIT Press.

Hansson, Gunnar O. (2001). *Theoretical and typological issues in consonant harmony*. PhD thesis, University of California, Berkeley.

Hansson, Gunnar O. and Ronald Sprouse (1999). Factors of change: Yowlumne vowel harmony then and now. In M. Caldecott, S. Gessner, and E.-S. Kim (Eds.), *Proceedings of WSCLA IV* (UBC Working Papers in Linguistics 2), pp. 39–57. Vancouver: Department of Linguistics, University of British Columbia.

Harris, J. W. (1969). *Spanish Phonology*. Cambridge, MA: MIT Press.

Harris, Zellig S. (1942). Morpheme alternants in linguistic analysis. *Language 18*, 169–180.

Harris, Zellig S. (1951). *Methods in Structural Linguistics*. The University of Chicago Press.

Harris, Z. S. (1955). From phoneme to morpheme. *Language 31*, 190–222. http://www.jstor.org/stable/411036

Harrison, K. D. and A. Kaun (2000). Pattern-responsive lexicon optimization. In M. Hirotani, A. Coetzee, N. Hall, and J. Kim (Eds.), *Proceedings of the Northeast Linguistics Society 30*. Amherst, MA: University of Massachusetts, Amherst, Graduate Linguistic Student Association. ROA 392.

Hay, Jennifer and Katie Drager (2007). Sociophonetics. *Annual Review of Anthropology 36*, 89–103.

Hay, Jennifer, Janet B. Pierrehumbert, and Mary Beckman. 2003. Speech perception, well-formedness, and the statistics of the lexicon. In *Papers in laboratory phonology VI*, ed. John Local, Richard Ogden, and Rosalind Temple, 58–74. Cambridge: Cambridge University Press.

Hayes, B. and C. Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39*, 379–440.

Hayes, Bruce, and Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23:59–104.

Hayes, Bruce, Robert Kirchner, and Donca Steriade, eds. 2004. *Phonetically-based phonology*. Cambridge: Cambridge University Press.

Hayes, B. (1995). On what to teach the undergraduates: Some changing orthodoxies in phonological theory. In I.-H. Lee (Ed.), *Linguistics in the Morning Calm, Volume 3*, pp. 59–77. Seoul: Hanshin.

Hayes, Bruce. 1999. Phonetically-driven phonology: the role of Optimality Theory and inductive grounding. In *Functionalism and formalism in linguistics, volume I: general papers*, ed.

Mike Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatley, 243– 285. Amsterdam: John Benjamins.

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

Heinz, J. (2007). *Inductive Learning of Phonotactic Patterns*. Ph. D. thesis, UCLA.

Heinz, Jeffrey. to appear-a. Learning phonotactic patterns from surface forms. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, ed. Donald Baumer, David Montero, and Michael Scanlon. Somerville, Mass.: Cascadilla Proceedings Project.

Heinz, Jeffrey. to appear-b. Learning quantity-insensitive stress patterns via local inference. In *Proceedings of The Association for Computational Linguistics Special Interest Group in Phonology 6 (ACL-SIGPHON 06)*. East Stroudsburg, Penn.: Association for Computational Linguistics.

Hock, H. H. (1991). *Principles of Historical Linguistics* (2nd ed.). Mouton de Gruyter.

Hockett, C. F. (1955). *A Manual of Phonology*. Baltimore: Waverly Press.

Hooper, J. (1976). Introduction to Natural Generative Phonology. New York: Academic Press.

Inkelas, S. (1995). The consequences of optimization for underspecification. In E. Buckley and S. Iatridou (Eds.), *Proceedings of the Twenty-Fifth Northeastern Linguistics Society*, pp. 287–302. Amherst: GLSA. ROA 40.

Inkelas, S. and C. Zoll (2003) Is Grammar Dependence Real? Ms, UC Berkeley and MIT. ROA 587.

Itô, J. and A. Mester (2002). The phonological lexicon. In N. Tsujimura (Ed.), *A Handbook of Japanese Linguistics*. Oxford: Blackwell.

Itô, J. and A. Mester (2003). On the sources of opacity in OT: coda processes in German. In C. Féry and R. van de Vijver (Eds.), *The Syllable in Optimality Theory*, pp. 271–303. Cambridge University Press

Jäger, Gerhard. 2004. Maximum entropy models and stochastic Optimality Theory. Rutgers Optimality Archive 625.

Jarosz, Gaja. 2006. Rich lexicons and restrictive grammars – maximum likelihood learning in Optimality Theory. Doctoral dissertation, Johns Hopkins University, Baltimore, Md.

Jarosz, G. (in press). Restrictiveness in phonological grammar and lexicon learning. In *Proceedings of the 43rd Annual Meeting of the Chicago Lingusitics Society*.

Jurafsky, D. and J. H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. NJ: Prentice Hall.

Jusczyk, Peter W., Angela D. Friederici, Jeanine M.I. Wessels, Vigdis Y. Svenkerud & Ann Marie Jusczyk (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* **32**: 402-420.

Jusczyk, Peter W., Paul A. Luce & Jan Charles-Luce (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* **33**: 630-645.

Jusczyk, P. W., P. Smolensky, and T. Allocco (2002). How English-learning infants respond to markedness and faithfulness constraints. *Language Acquisition 10*, 31–73.

Kager, R. (1999). Optimality Theory. Cambridge University Press.

Kager, R. (in press). Lexical irregularity and the typology of contrast. In K. Hanson and S. Inkelas (Eds.), The Nature of the Word–Essays in Honor of Paul Kiparsky. MIT Press.

Kaisse, Ellen M. (1986). Locating Turkish devoicing. In *WCCFL 5*, pp. 119–128.

Kawahara, Shigeto. 2002. Similarity among Variants: Output-Variant Correspondence. Bachelor's thesis, International Christian University.

Kazazis, K. (1969). Possible evidence for (near-)underlying forms in the speech of a child. *Chicago Linguistics Society 5*, 382–386.

Keating, P. A. (1985). Universal phonetics and the organization of grammars. In V. Fromkin (Ed.), *Phonetic Linguistics*, pp. 115–132. Academic Press.

Kemler Nelson, D. G., P. W. Jusczyk, D. R. Mandel, J. Myers, A. Turk, and L. A. Gerken (1995) The Headturn Preference Procedure for Testing Auditory Perception. *Infant Behavior and Development* 18, 111–116.

Kenstowicz, Michael (1997). Base identity and uniform exponence: Alternatives to cyclicity. In J. Durand and B. Laks (Eds.), *Current Trends in Phonology: Models and Methods*, pp. 363–394. Salford: University of Salford.

Kerkhoff, A. (2007). Acquisition of Morpho-Phonology: The Dutch voicing alternation. Ph.D. thesis, University of Utrecht.

King, Robert D. (1969). *Historical Linguistics and Generative Grammar*. Englewood Cliffs, N.J.: Prentice-Hall.

King, Robert D. (1973). Rule insertion. *Language* 49, 551–578.

King, Robert D. (1980). The history of final devoicing in Yiddish. In M. I. Herzog, B. Kirshenblatt-Gimblett, D. Miron, and R. Wisse (Eds.), *The Field of Yiddish: Studies in Language, Folklore, and Literature, Fourth Collection*, pp. 371–430. Philadelphia: Institute for the Study of Human Issues.

Kingston, J. and R. L. Diehl (1994). Phonetic knowledge. *Language 70*, 419–454.

Kiparsky, P. (1965). Phonological Change. Ph. D. thesis, MIT.

Kiparsky, P. (1971). Historical linguistics. In W. O. Dingwall and W. O. Dingwall (Eds.), *A Survey of Linguistic Science*, pp. 576–642. College Park, MD: University of Maryland Linguistics Program.

Kiparsky, P. (1972). Explanation in phonology. In S. Peters (Ed.), Goals in Linguistic Theory, pp. 189–227. Englewood Cliffs, NJ: Prentice-Hall.

Kiparsky, P. and L. Menn (1977). On the acquisition of phonology. In J. Macnamara (Ed.), *Language learning and thought*. New York: Academic Press.

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

Kuroda, S.-Y. (1967). *Yawelmani Phonology*. Cambridge, Mass.: MIT Press.

Legendre, G., Y. Miyata, and P. Smolensky. 1990. *Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations*. Technical Report 90-5, Institute of Cognitive Science, Univ. of Colorado.

Lin, Ying. 2005a. Learning Stochastic OT Grammars: A Bayesian approach using Data Augmentation and Gibbs Sampling. *Proceedings of the Association for Computational Linguistics*. East Stroudsburg, Penn.: Association for Computational Linguistics.

Lin, Ying. 2005b. Learning features and segments from waveforms: a statistical model of early phonological acquisition. Doctoral dissertation, University of California, Los Angeles.

Lin, Ying. Ms. Stochastic Optimality Theory, local search, and Bayesian learning of hierarchical linguistic models. Ms., Department of Linguistics, University of Arizona. Downloaded from http://dingo.sbs.arizona.edu/%7Eyinglin/Lin_hierarchical.pdf

Ling, C. and M. Marinov (1993). Answering the connectionist challenge: a symbolic model of learning the past tenses of English verbs. *Cognition* 49, 235–290.

Manuel, S. Y. and R. A. Krakow (1984). Universal and language particular aspects of vowel-to-vowel coarticulation. *Haskins Laboratories Status Reports on Speech Research SR-77/78*, 69–78.

MacBride, A. (2004). A Constraint-Based Approach to Morphology. Ph. D. thesis, UCLA.

MacEachern, Margaret R. 1999. Laryngeal Cooccurrence Restrictions. Garland Publishing.

MacWhinney, Brian. 1975. Rules, rote, and analogy in morphological formations by Hungarian children. *Journal of Child Language* 2: 65–77.

Maddieson, Ian (1985). Phonetic cues to syllabification. *Phonetic Linguistics*, 203–221.

Maiden, M. (2005). Morphological autonomy and diachrony. In G. Booij and J. van Marle (Eds.), *Yearbook of Morphology 2004*, pp. 137–175. Springer.

Maslova, Elena. To appear. Stochastic OT as a model of constraint interaction. To appear in Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson, and Annie Zaenen (eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. CSLI publications.

Matthews, P. H. (1972). *Inflectional morphology: a theoretical study based on aspects of Latin verb conjugation*. Cambridge: Cambridge University Press.

Maye, J., J. Werker, and L. Gerken (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition 82*, B101–B111.

McCarthy, J. J. (2001). *A Thematic Guide to Optimality Theory*. Cambridge University Press.

McCarthy, J. (1998). Morpheme structure constraints and paradigm occultation. In M. C. Gruber, D. Higgins, K. Olson, and T. Wysocki (Eds.), *CLS 32, vol. II: The Panels*, Chicago. Chicago Linguistic Society.

McCarthy, J. (2005). Taking a free ride in morphononemic learning. *Catalan Journal of Linguistics* 4, pp. 19–55. ROA 683.

McCarthy, John. (2007) *Hidden Generalizations: Phonological Opacity in Optimality Theory*. London: Equinox.

McCawley, J. (1967). Sapir's phonologic representation. *International Journal of American Linguistics* 33(2), 106–111.

Merchant, Nazarré and Bruce Tesar (2008). Learning underlying forms by searching restricted lexical subspaces. In *Proceedings of the Forty-First Conference of the Chicago Linguistics Society (2005), vol. II*, pp. 33–48. Chicago: Chicago Linguistics Society.

Mielke, Jeff. 2004. The emergence of distinctive features. Doctoral dissertation, The Ohio State University.

Mielke, J. (2005). Modeling distinctive feature emergence. In J. Alderete, C. Han, and A. Kochetov (Eds.), *Proceedings of the 24th West Coast Conference on Formal Linguistics*, pp. 281–289. Somerville, MA: Cascadilla Proceedings Project.

Mielke, J., M. Armstrong, and E. Hume (2003). Looking through opacity. *Theoretical Linguistics* 29, 123–139.

Mielke, Jeff. 2005. Ambivalence and ambiguity in laterals and nasals. *Phonology* 22: 169–203.

Mikheev, Andrei. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics* 23:405–423.

Moreton, Elliott (2008). Analytic bias and phonological typology. *Phonology 25*, 83–127.

Nagy, Naomi and Bill Reynolds. 1997. Optimality Theory and word-final deletion in Faetar. *Language Variation and Change* 9:37–55.

Nakisa, R. C., K. Plunkett, and U. Hahn (1997). A Cross-Linguistic Comparison of Single and Dual-Route Models of Inflectional Morphology. In P. Broeder and J. Murre (Eds.), *Cognitive Models of Language Acquisition*. Cambridge, MA: MIT Press.

Newport, Elissa L. and Richard N. Aslin. 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology* 48:127–162.

Niyogi, Partha and Robert Berwick (1996) A language learning model for finite parameter spaces. In Michael R. Brent, ed., Computational approaches to language acquisition, 605–622. Cambridge, MA: MIT Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115, 39–57.

Nosofsky, R. M. (1990). Relations between exemplar similarity and likelihood models of classification. *Journal of Mathematical Psychology* 34, 393–418.

Ohala, John J. 1981. The listener as the source of sound change. In *Papers from the Parasession on Language and Behavior*, ed. Carrie S. Masek, Roberta. A. Hendrick, and Mary Frances Miller, 178–203. Chicago, Illinois: Chicago Linguistic Society.

Ohala, John J., and Manjari Ohala. 1986. Testing hypotheses regarding the psychological reality of morpheme structure constraints. In *Experimental phonology*, ed. John J. Ohala and Jeri J. Jaeger, 239–252. San Diego, Calif.: Academic Press.

Oncina, J., P. García, and E. Vidal (1993). Learning subsequential transducers for pattern recognition interpretation tasks. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 448–458.

Pater, Joe (2000) Nonuniformity in English stress: The role of ranked and lexically specific constraints. *Phonology 17*, 237–274.

Pater, Joe (2004). Bridging the gap between perception and production with minimally violable constraints. In R. Kager, J. Pater, , and W. Zonneveld (Eds.), *Constraints in Phonological Acquisition*, pp. 219–244. Cambridge University Press.

Pater, Joe (2008a). Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In Steve Parker (ed.) *Phonological Argumentation: Essays on Evidence and Motivation*, Equinox. 1–33.

Pater, Joe. (2008b) Gradual learning and convergence. *Linguistic Inquiry* 39. 334-345.

Pater, Joe, Rajesh Bhatt and Christopher Potts. 2007. Linguistic Optimization. Ms, University of Massachusetts, Amherst.

Pater, Joe, Christopher Potts, and Rajesh Bhatt. 2006. Harmonic grammar with linear programming. Rutgers Optimality Archive 872.

Pater, Joe, and Anne-Michelle Tessier. 2003. Phonotactic knowledge and the acquisition of Alternations. Proceedings of the 15th International Congress on Phonetic Sciences, ed. by María-Josep Solé, Daniel Recasens, and Joaquín Romero, 1177-1180. Barcelona: Casual Productions.

Pater, Joe, and Anne-Michelle Tessier. 2006. L1 phonotactic knowledge and the L2 acquisition of alternations. Inquiries in linguistic development: Studies in honor of Lydia White, ed. by R. Slabakova, S. Montrul, and P. Prévost, 115-131. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. (2006). The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition* B31-B41.

Pertz, D. L. and T. G. Bever (1975). Sensitivity to phonological universals in children and adolescents. *Language 51*, 149–162.

Pierrehumbert, Janet. 1994. Syllable structure and word structure: a study of triconsonantal clusters in English. In *Phonological structure and phonetic form: papers in laboratory phonology III*, ed. Patricia Keating, 168–188. Cambridge: Cambridge University Press.

Pierrehumbert, J. B. (2006). The statistical basis of an unnatural alternation. In L. M. Goldstein, D. H. Whalen, and C. T. Best (Eds.), *Laboratory Phonology VIII: Varieties of Phonological Competence*, pp. 81—106. Berlin: Mouton de Gruyter.

Pinker, S. and A. Prince (1988). On language and connectionism: Analysis of a Parallel Distributed Processing model of language acquisition. *Cognition 28*, 73–193.

Poliquin, Gabriel (2006). *Canadian French Vowel Harmony*. Ph. D. thesis, Harvard University.

Prince, A. (1983). Relating to the grid. Linguistic Inquiry 4, 19–100.

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Fixing priorities: constraints in phonological acquisition*, ed. René Kager, Joe Pater, and Wim Zonneveld, 245– 291. Cambridge: Cambridge University Press.

Prince, Alan, and Paul Smolensky. 1993/2004. *Optimality Theory: constraint interaction in generative grammar*. Cambridge, Mass.: Blackwell. [Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, N.J., April 1993].

Reynolds, William. 1994. Variation and phonological theory. Doctoral dissertation, University of Pennsylvania, Philadelphia.

Rice, Keren. 2005. Liquid relationships. In Chiara Frigeni, Manami Hirayama, Sara Mackenzie (eds.) Toronto Working Papers in Linguistics, Volume 24: Special Issue on Similarity in Phonology, 31–44.

Riggle, Jason. 2004. Generation, recognition, and learning in finite state Optimality Theory. Doctoral dissertation, University of California, Los Angeles.

Riggle, J. (2006). Using entropy to learn ot grammars from surface forms alone. In D. Baumer, D. Montero, and M. Scanlon (Eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pp. 346–353. Somerville, MA: Cascadilla Proceedings Project.

Russell, Kevin (1999). MOT: Sketch of an ot approach to morphology. ROA 352.

Sanders, N. (2003). *Opacity and sound change in the Polish lexicon*. Ph. D. thesis, University of California, Santa Cruz. ROA 603

Scholes, Robert. 1966. *Phonotactic grammaticality*. The Hague: Mouton.

Schütze, C. (2005). Thinking about what we are asking speakers to do. In S. Kepser and M. Reis (Eds.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, pp. 457–484. Mouton de Gruyter.

Sievers, Eduard. 1901. *Grundzüge der Phonetik*, 5th ed. Leipzig: Breitkopf und Härtel.

Skousen, R. (1989). *Analogical Modeling of Language.* Dordrecht: Kluwer Academic Publishers.

Smith, C. (1992). The temporal organization of vowels and consonants. Ph. D. thesis, Yale University.

Smith, Jennifer L. 2000. Positional faithfulness and learnability in Optimality Theory. In Rebecca Daly and Anastasia Riehl, eds., *Proceedings of ESCOL '99*. Ithaca, NY: CLC Publications, 203-214.

Smith, Jennifer L. 2004. Making constraints positional: Toward a compositional model of CON. *Lingua* 114: 1433–1464.

Smith, Neilson V. (1973) *The acquisition of phonology: A case study*. Cambridge: Cambridge University Press.

Smolensky, P. (1996). The initial state and 'Richness of the Base' in Optimality Theory. Technical report, Johns Hopkins University.

Smolensky, P. (1996). On the comprehension/production dilemma in child language. Linguistic Inquiry 27, 720–731.

Smolensky, Paul, and Géraldine Legendre. 2006. *The harmonic mind: from neural computation to Optimality-theoretic grammar*. Cambridge, Mass.: MIT Press.

Sommerstein, Alan H. 1974. On phonotactically motivated rules. *Journal of Linguistics* 19:71–94.

Steriade, Donca. 1999. Alternatives to syllable-based accounts of consonantal phonotactics. In *Proceedings of the 1998 Linguistics and Phonetics Conference*, ed. Osamu Fujimura, Brian Joseph, and B. Palek, 205–245. Prague: The Karolinum Press.

Steriade, Donca (2000). Paradigm uniformity and the phonetics/phonology boundary. In M. Broe and J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge University Press.

Steriade, Donca. In press. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. In A *Festschrift for Paul Kiparsky*.

Stump, G. T. (2001). *Inflectional morphology: A theory of paradigm structure*. Cambridge: Cambridge University Press.

Tesar, B., J. Alderete, G. Horwood, N. Merchant, K. Nishitani, and A. Prince. 2003. Surgery in language learning. In G. Garding and M. Tsujimura, eds. *WCCFL 22*, 477–490. Somerville, MA: Cascadilla Press. ROA-619.

Tesar, Bruce, and Alan Prince. 2003. Using phonotactics to learn phonological alternations. In *Proceedings of the Thirty-Ninth Conference of the Chicago Linguistics Society, vol. II: The panels*. Chicago: Chicago Linguistic Society.

Tesar, B. and A. Prince (2007). Using phonotactics to learn phonological alternations. In J. E. Cihlar, A. L. Franklin, D. W. Kaiser, and I. Kimbara (Eds*.), CLS 39–2: The Panels: Papers from the 39th Annual Meeting of the Chicago Linguistic Society*, pp. 209–237. Chicago Linguistic Society.

Tesar, Bruce, and Paul Smolensky. 1993. The learnability of Optimality Theory: An algorithm and some basic complexity results. Ms. Department of Computer Science and Institute of Cognitive Science, University of Colorado at Boulder. Rutgers Optimality Archive ROA-2, http://ruccs.rutgers.edu/roa.html.

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, Mass.: MIT Press.

Tesar, Bruce. 1995. Computational Optimality Theory. Doctoral dissertation, University of Colorado.

Tesar, Bruce. 2004. Using inconsistency detection to overcome structural ambiguity. *Linguistic Inquiry* 35:219–253.

Tessier, Anne-Michelle. 2006. Biases and Stages in Phonological Acquisition. PhD thesis, University of Massachusetts Amherst.

Treiman, Rebecca, Brett Kessler, Stephanie Knewasser, Ruth Tincoff, and Margo Bowman. 2000. English speakers' sensitivity to phonotactic patterns. In *Papers in laboratory* *phonology v: acquisition and the lexicon*, ed. Michael B. Broe and Janet Pierrehumbert, 269–282. Cambridge: Cambridge University Press.

Tuller, B. and J. Kelso (1991). The production and perception of syllable structure. *Journal of Speech and Hearing Research 34*, 501–508.

Vennemann, T. (1972). Phonetic analogy and conceptual analogy. In T. Vennemann and T. H. Wilbur (Eds.), *Schuchhardt, the Neogrammarians, and the Transformational Theory of Phonological Change: Four Essays by Hugo Schuchhardt*, Number 26 in Linguistische Forschungen, pp. 115–179. Frankfurt am Main: Athenäum.

Vennemann, T. Phonological concreteness in Natural Generative Grammar. In R. Shuy and C.-J. N. Bailey (Eds.), *Toward Tomorrow's Linguistics*. Washington D.C.: Georgetown University Press.

Vitevitch, M., Luce, P., Charles-Luce, J., and Kemmerer, D. (1996). Phonotactic and metrical influences on adult ratings of spoken nonsense words. In *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA.

Wilson, Colin. 2006. Learning phonology with substantive bias: an experimental and computational investigation of velar palatalization. *Cognitive Science* 30:945–982

Wilson, Colin. Ms. The Luce choice ranker. Ms., Department of Linguistics, UCLA.

Wurzel, W. U. (1987). Paradigmenstrukturbedingungen: Aufbau und Veränderung von Flexionsparadigmen. In A. G. Ramat, O. Carruba, and G. Bernini (Eds.), *Papers from the 7th International Conference on Historical Linguistics*, pp. 629–644. John Benjamins Publishing Company.

Yu, Alan C. L. (2004). Explaining final obstruent voicing in Lezgian: Phonetics and history. Language 80, 73–97.

Zamuner, Tania S., Annemarie Kerkhoff, and Paula Fikkert (2006). Acquisition of voicing neutralization and alternations in Dutch. In *Proceedings of the 30th Boston University Conference on Language Development*, ed. David Bamman, Tatiana Magnitskaia, and Colleen Zaller, 701–712. Somerville, MA: Cascadilla Press.

Zhang, Jie, Yuwen Lai, and Craig Turnbull-Sailor (2006). Wug-testing the "tone circle" in Taiwanese. In Donald Baumer, David Montero, and Michael Scanlon (eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics*. Cascadilla Proceedings Project, Somerville, MA, pp. 453–461.

Zuraw, Kie (2000). *Patterned Exceptions in Phonology*. Ph. D. thesis, UCLA.

Zwicky, A. M. (1985). How to describe inflection. In M. Niepokuj, M. VanClay, V. Nikiforidou, and D. Feder (Eds.), Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society, February 16-18, 1985, pp. 372–386. Berkeley Linguistics Society.