ELSEVIER

Lingua 104 (1998) 131–145

Lingua

An iterative strategy for language learning☆

Bruce Tesar

Rutgers Center for Cognitive Science, Linguistics Department, Rutgers University, Piscataway, NJ 08855, USA

Abstract

One of the major challenges of language acquisition is the fact that the auditory signal received by a child underdetermines the structural description of the utterance. This paper approaches the problem by capitalizing on the optimizing structure of Optimality Theory. The learner uses a hypothesized grammar to make a best guess at the full structural description of an observed overt form, filling in the hidden structure not apparent. The learner uses it to modify their grammar, despite the fact that the full description is based in part on the previous (most likely wrong) grammar. The claim is that the learner can go back and forth between estimating the hidden structure and estimating the grammar, eventually converging on the correct grammar. The results of some simulations are presented in support of the claim.

1. The problem of learning hidden structure

A central challenge of language acquisition is the indirectness of the relationship between the overt, auditory information accessible by the learner, and the space of possible grammars provided by universal grammar. One source of indirectness is the fact that an *overt form* (the overt information of an uttered linguistic form) can underdetermine the structural description of that form assigned by the target grammar. To use metrical stress as an example, consider an overt form consisting of three light syllables, with a single stress on the middle syllable: [0 1 0]. There are two distinct structural descriptions consistent with this overt form: one with an iambic left-aligned foot and the final syllable unfooted, [(0 1) 0], and one with a trochaic right-aligned foot and the initial syllable unfooted, [0 (1 0)]. The overt form by itself does not provide a basis for distinguishing between the two; the foot structure itself is 'hidden', and the overt form is ambiguous. This is not parochial to metrical stress; it could be argued that all areas of linguistics postulate representational structures not

☆ tesar@ruccs.rutgers.edu

* The author would like to thank Alan Prince and Paul Smolensky for many valuable discussions. Helpful comments and discussion were also provided by Elan Dresher, John McCarthy, Joe Pater, two anonymous *Lingua* reviewers, the participants of the second Rutgers/UMass Joint Class Meeting, and especially the organizers and participants of the BCN Workshop on Conflicting Constraints.

PII S0024-3841(97)00027-2

directly apparent from overt forms, and therefore create the possibility that some overt forms are ambiguous. Further, the hidden structure in the full structural descriptions are not just illustrative devices for the convenience of description; the explanatory principles of linguistic theory make crucial reference to the hidden structure.

Indirectness between overt forms and grammars also results from the fact that the same structural description may be licensed by more than one possible grammar, and for different reasons. For example, the description [(1 0) 0] might be the result of a requirement that the head foot be aligned with the left edge of the word, or it might be the result of extrametricality of the final syllable. The description itself does not give an answer; other descriptions, such as [(1 0) (2 0)] (former case) or [0 (1 0) 0] (latter case), taken together with this one, are needed to determine the particular grammar at work. Clearly, this problem is related to the first; the relationship between the grammars and the descriptions is dependent upon the 'hidden' elements of the descriptions (here, the foot structure).

Much recent work in language learnability (Dresher and Kaye, 1990; Clark and Roberts, 1993; Gibson and Wexler, 1994) has treated the indirectness of the relationship between overt forms and grammars in monolithic terms, without identifying the mediating role of full structural descriptions.[1] However, Tesar and Smolensky have suggested that decomposing the learning problem along these lines can be beneficial (Tesar, 1998; Tesar and Smolensky, 1996). In support of this, they presented the *constraint demotion* procedure, which completely solves the subproblem of the relationship between full structural descriptions and grammars. Their solution, described in section 3.2, does not come for free; it applies to the framework of Optimality Theory (Prince and Smolensky, 1993), and crucially relies on the commitment that cross-linguistic variation is accounted for entirely by different rankings of the same universal constraints. As a solution to this subproblem, constraint demotion has several appealing properties: (a) it applies to all linguistic analyses within the OT framework, not just those of, say, stress, or even phonology; (b) it is guaranteed to find a correct ranking; and (c) it is quite fast.

This leaves, however, the other subproblem, that of the underdetermination of full structural descriptions by overt forms. Tesar and Smolensky outlined a proposal for approaching this problem, an approach that would not require abandoning all of the benefits of constraint demotion. This paper presents results of an initial investigation into that proposal. The learner uses the optimizing structure of Optimality Theory to estimate the hidden structure associated with an overt form, thus providing a hypothesized full structural description. That hypothesized description is then used to modify the learner's constraint ranking via constraint demotion. What results is an iterative procedure, one which alternates between using a constraint ranking to hypothesize hidden structure, and using the hypothesized hidden structure to refine the ranking. Success occurs when this procedure converges on a correct ranking, as indicated by consistency between the ranking and all available overt forms.

---

[1] It is in this regard that Dresher's characterization of learning in terms of the Credit Problem and the Epistemological Problem (Dresher, 1996) differ from the characterization given here.

2. An optimality-theoretic system of stress grammars

Metrical stress theory has been a domain of focus for several learning investigations. Dresher and Kaye (1990) applied cue learning to a system of stress grammars set within the principles and parameters framework (Chomsky, 1981). Approaches less closely tied to any explicit linguistic framework have also been investigated (Gupta and Touretzky, 1994; Daelemans et al., 1994). Metrical stress is an appealing domain because a lot is known about it, and because it can be treated somewhat in isolation from other aspects of phonology.

Metrical stress was selected for the current investigation because it permits the issue of input/output faithfulness to be set aside. In the present analysis, underlying forms are strings of syllables, and structural descriptions assign stresses to the syllables; no insertion/deletion of syllables is considered (for discussion of learning underlying forms, including relations with child language acquisition work, see Tesar and Smolensky (1996), Smolensky (1996), and the works cited therein). Thus, the underlying form for an utterance can be directly (and correctly) inferred from the overt form; the underlying form is simply the syllables of the overt form (without the stresses). The following optimality-theoretic system for metrical stress is loosely based upon analyses developed by McCarthy and Prince (1993). It captures a subset of attested metrical phenomena: it includes main-stress only and iterative footing systems, as well as effects traditionally analyzed as directional footing and extrametricality. It does not include any analysis of quantity sensitivity effects.

Each structural description is of a single prosodic word, and all overt forms are of single prosodic words. The overt forms are strings of stress levels, one for each syllable. The overt forms range from 2 to 7 syllables in length. A structural description is a grouping of the syllables (with their stress levels) into feet. Table 1 shows the pairings of overt forms and structural descriptions for the stress pattern of Garawa (Furby, 1974; Hayes, 1995); the analysis is taken from McCarthy and Prince (1993). Prosodic word boundaries are denoted as square brackets, and foot boundaries are denoted as parentheses. The GEN function will only generate descriptions in which each foot has precisely one head syllable, which is the sole stress-bearing syllable of that foot. An unfooted syllable must be unstressed. GEN also requires that a prosodic word have precisely one head foot, whose head syllable bears main stress. If the word has any other (non-head) feet, their head syllables each bear secondary stress. Feet are strictly bisyllabic.

Table 1
The Garawa Stress Pattern (1 = main stress, 2 = secondary stress, 0 = unstressed)

Overt forms	Descriptions
[1 0]	[(1 0)]
[1 0 0]	[(1 0) 0]
[1 0 2 0]	[(1 0) (2 0)]
[1 0 0 2 0]	[(1 0) 0 (2 0)]
[1 0 2 0 2 0]	[(1 0) (2 0) (2 0)]
[1 0 0 2 0 2 0]	[(1 0) 0 (2 0) (2 0)]

The system has 11 constraints, listed in Table 2. The 11 constraints are freely rankable. The PARSE-SYLLABLE constraint is violated by any unfooted syllable; its ranking determines when and if secondary stresses will occur. The ALL-FEET-RIGHT/LEFT constraints pressure all feet to be as close to the right/left edge of the word as possible, and help capture phenomena previously analyzed as directional iterativity. A sufficiently high ranking of NON-INITIAL and/or NON-FINAL produces syllable extrametricality effects (Liberman and Prince, 1977; Hayes, 1980) (the formulation of NON-FINAL given here is closer to the traditional conception of extrametricality than the formulation given in Prince and Smolensky, 1993). Four of the constraints, MAIN-RIGHT/LEFT and ALL-FEET-RIGHT/LEFT, are gradient alignment constraints between edges of feet and the edges of the prosodic word. Each such constraint assesses a constraint violation for every syllable intervening between the relevant foot-edge and the relevant word-edge.

Table 2
The constraints

Name	Description
PARSE-SYLLABLE	a syllable must be footed
MAIN-RIGHT	align the head-foot with the word, on the right edge
MAIN-LEFT	align the head-foot with the word, on the left edge
ALL-FEET-RIGHT	align each foot with the word, on the right edge
ALL-FEET-LEFT	align each foot with the word, on the left edge
WORD-FOOT-RIGHT	align the right edge of the word with some foot
WORD-FOOT-LEFT	align the left edge of the word with some foot
IAMBIC	align the head syllable with its foot, on the right edge
TROCHAIC	align the head syllable with its foot, on the left edge
NON-FINAL	the final syllable should not be footed
NON-INITIAL	the initial syllable should not be footed

Table 3 shows a constraint hierarchy which generates the stress pattern for Garawa shown in Table 1. The four constraints in the top stratum are effectively undominated for this language: any relative ranking among them will produce the same language, so long as all four of them dominate all of the remaining constraints, and the other constraints are held in the same position relative to each other. High-ranked TROCHAIC, in virtue of dominating IAMBIC, ensures that each foot is trochaic; a stressed syllable is aligned with the left edge of its foot. MAIN-LEFT determines that the head foot of the prosodic word (assigning main stress) is at the left edge of the word. These two constraints, ranked at the top, ensure that main stress always falls on the initial syllable. Top-ranked PARSE-SYLLABLE ensures that words with four or more syllables will have additional feet assigning secondary stresses. The constraint in the second stratum, ALL-FEET-RIGHT, dominates all of the remaining constraints. It applies to each foot, and for each foot is violated by each syllable separating the foot from the right edge of the prosodic word. Obviously, at most one foot can be perfectly aligned with the right edge of the word; if there are multiple feet, some violations of the constraint are inevitable. However, the number of violations is

Table 3
A constraint ranking generating the Garawa stress pattern

{PARSE-SYLLABLE WORD-FOOT-LEFT TROCHAIC MAIN-LEFT}
{ALL-FEET-RIGHT}
{NON-INITIAL NON-FINAL IAMBIC WORD-FOOT-RIGHT ALL-FEET-LEFT MAIN-RIGHT}

reduced if the feet are stacked up towards the right. The effect in Garawa is seen in 5- and 7-syllable words. In those words, the feet align from the right, except the head foot, which stays aligned to the left edge of the word, incurring an extra violation of ALL-FEET-RIGHT rather than a violation of higher-ranked MAIN-LEFT. Because MAIN-LEFT only applies to the head foot, it has no effect on the other feet of the word, permitting ALL-FEET-RIGHT to determine their position.

Under the formal definition of Optimality Theory, a grammar requires a *total ranking* of the constraints, with the relative ranking determined for every pair of constraints. However, the learning algorithm makes use of a more general space of hypotheses, that of stratified hierarchies (see Tesar, 1998, for discussion of the role of stratified hierarchies in learning). In a stratified hierarchy, one or more constraints may occupy the same stratum in a hierarchy. The constraints of a stratum are not ranked relative to each other, but all of them dominate all constraints occupying lower strata. The use of stratified hierarchies requires that the definition of an OT mapping be extended. Two candidates are compared on a stratified hierarchy as follows. The candidates are evaluated on all constraints in the top stratum, and it is determined, for each constraint, which candidate has more violations. The candidate which has greater violation of fewer of the constraints in the stratum is the better one. A constraint violated equally by both candidates makes no contribution to distinguishing the two (as always). If the two candidates fare equally on the top stratum, the decision is passed to the next stratum, and so forth (this extension is slightly different from the one given in Tesar and Smolensky, 1996). The typological predictions of the theory are unchanged; a stratified hierarchy only counts as a target grammar if it generates a language also generated by at least one total ranking of the constraints (as is the case for the stratified hierarchy in Table 3).

3. The iterative learning strategy

3.1. Optimization models of language processing

In Optimality Theory, the constraint ranking determines which of the many candidate structural descriptions is optimal (hence grammatical) for a given underlying form. This mapping provides a (highly idealized) characterization of language production as an optimization process. The process of computing the optimal structural description for an underlying form, given a constraint ranking, will be called *production-directed parsing*. The algorithms used for parsing in these simulations are based on the parsing algorithms developed by Tesar (1995). Other work on efficient

algorithms for production-directed parsing with OT grammars has been done by T. Mark Ellison (1994).

A corresponding mapping for language comprehension would be one which maps from an overt form of the language to its grammatical structural description. A definition for such a mapping was suggested by Tesar and Smolensky (1996): the hearer is presented with an overt form, and selects the structural description of that overt form that is optimal with respect to their current constraint ranking. The difference is that here the candidate structural descriptions competing for optimality are candidates whose overt portions match the observed overt form. In principle, this means that structural descriptions for different underlying forms might compete with each other, if they have identical overt forms. A structural description with an overt form that matches the observed overt form is an *interpretation* of that overt form. The interpretation assigned to an overt form is that structural description which, out of all descriptions whose overt portion matches the overt form, best satisfies the ranked universal constraints. The process of computing the optimal interpretation for an overt form, given a constraint ranking, will be called *interpretive parsing*. Other work, within Optimality Theory, on parsing overt forms, specifically elements of syllable structure, has been done by Hammond (1995).

Language production and language comprehension, then, are both optimization processes. When a competent speaker, possessing the correct constraint ranking, interprets an overt form, they arrive at the same structural description as when they apply production-directed processing to the corresponding underlying form. Both processes are defined in terms of optimization with respect to the same constraint ranking. What is proposed here is that a language learner uses the same processes during learning. At any given time, the learner has a ranking; this ranking is supplied to production-directed parsing when attempting to produce language, and it is supplied to interpretive parsing when the learner is interpreting the overt forms it hears. (See Fig. 1.)

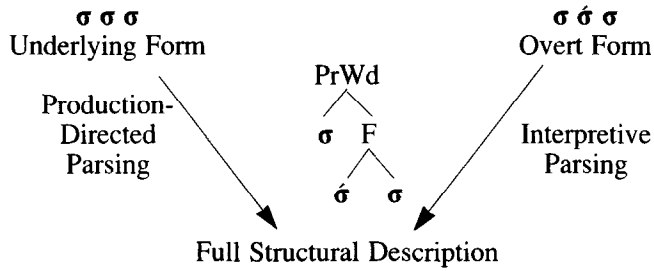

Fig. 1. The relationship between production-directed and interpretive parsing

One critical property of interpretive parsing is that it is 'robust' in the following sense: the process assigns a description to an overt form even when there is no description matching that overt form which is grammatical (i.e., optimal for its underlying form) according to the current ranking. A description is still assigned to the overt form, even if there is another description for the same underlying form (but

with a different overt form) which better satisfies the ranked constraints. The learner can be aware that the utterance is not grammatical according to their current ranking, but nevertheless do the best they can to interpret the utterance. This reflects the observation that competent speakers can often offer consistent interpretations of utterances they simultaneously judge to be ungrammatical.

To illustrate this, consider the following example. The target language that a learner is exposed to is one with no secondary stress (i.e., main stress only), with a trochaic head foot aligned with the right edge of the prosodic word (giving penultimate main stress). However, the learner starts with a hypothesis ranking which assigns an iambic head foot aligned from the left edge of the word (giving peninitial main stress), and no secondary stress. The important part of the ranking is shown in equation (1) (for the purposes of space, the constraint names will be abbreviated):

A-F-L ≫ A-F-R ≫ IAMB ≫ MAIN-R ≫ TROCH ≫ 'the rest' (1)

Consider the overt form [0 0 0 1 0]. The learner, using the ranking in equation (1), applies interpretive parsing to arrive at the analysis of [0 0 (0 1) 0]; this is the most harmonic structural description consistent with the overt form. The learner then applies production-directed parsing to the underlying form of five syllables, getting [(0 1) 0 0 0]. This pair, along with their violations of the relevant constraints, is shown in Table 4. The optimal interpretation, [0 0 (0 1) 0], fares better on the current ranking than the other candidate consistent with the overt form, [0 0 0 (1 0)], because it has fewer violations of ALL-FEET-LEFT. The result of production-directed parsing, [(0 1) 0 0 0], fares better on the current ranking than the interpretive parse because it incurs fewer violations yet of ALL-FEET-LEFT.

Table 4
The production-directed parse of 5 syllables better satisfies the ranked constraints than the optimal interpretation of overt form [0 0 0 1 0], which in turn does better than an alternative interpretation

		A-F-L	A-F-R	IAMB	MAIN-R	TROCH
Production	[(0 1) 0 0 0]		* * *		* * *	*
Interpretation	[0 0 (0 1) 0]	* *	*		*	*
Alternate Interp.	[0 0 0 (1 0)]	* * *		*		

From the standpoint of learning, the most important fact about the production-directed parse and the optimal interpretation is that they are not identical: the observed word was not stressed in the same way that the learner's current ranking stresses it. This mismatch causes the learner to revise their ranking hypothesis, in an effort to find a ranking which will stress the word in the same way as was observed. More precisely, the learner will use their interpretation of the surface form, [0 0 (0 1) 0], as the target, modifying the ranking in an effort to make this description the optimal one for an underlying form of 5 syllables.

An additional point worth noticing: the learner has mis-analyzed the overt form; the correct analysis (the one assigned by the target grammar) is [0 0 0 (1 0)], the

other analysis consistent with the overt form. This might be cause for concern. However, as will be illustrated, it is sometimes possible for the learner to learn successfully even when an overt form is initially misinterpreted.

3.2. Constraint demotion

Error-Driven Constraint Demotion is a procedure for learning constraint rankings on the basis of grammatical full structural descriptions (including both overt and hidden structure). The procedure works by pairing the target description (one that is grammatical in the target language), called the *winner*, with the description for the same underlying form that is optimal under the learner's current ranking, called the *loser* (computed by production-directed parsing). The procedure then changes the ranking to one better satisfied by the winner than by the loser. The method of changing the constraint ranking gives constraint demotion its name. The constraints violated more times by the winner are demoted to below the highest constraint violated more times by loser; this ensures that, in the new ranking, the loser violates (more times) a constraint dominating all constraints violated (more times) by the winner. If a constraint violated more times by the winner is already dominated by a constraint violated more by the loser, it is left alone. An important property of constraint demotion is that, when provided with the correct full descriptions of a language, it is guaranteed to find a correct ranking of the constraints, and quite efficiently. See Tesar and Smolensky (1996) and Tesar (1998): for further discussion, including proofs of correctness and time bounds.

To continue the example of section 3.1, the learner wants to modify the ranking so that the interpretive parse, [0 0 (0 1) 0], does better than the production-directed parse. Thus, [0 0 (0 1) 0] is the winner, [(0 1) 0 0 0] the loser. The highest-ranked constraint assessing more marks to the loser is ALL-FEET-RIGHT, so the constraint assessing more marks to the winner, ALL-FEET-LEFT, is demoted to immediately below ALL-FEET-RIGHT, into the stratum already occupied by IAMBIC; this is shown in Table 5. The loser now loses to the winner, because it has more violations of ALL-FEET-RIGHT. While this demotion ensures that the winner beats this loser it does not ensure that the optimal description matches the observed overt form. Now that the learner has a new ranking, it must go back to the beginning of the procedure, using the new ranking. The rest of this example is given in section 3.3.

3.3. Iterating learning and parsing

The learner began by applying interpretive parsing to an observed overt form, [0 0 0 1 0], getting interpretation [0 0 (0 1) 0]. It then applied production-directed pars-

Table 5
After the first demotion, the winner is more harmonic

		A-F-R	IAMB	A-F-L	MAIN-R	TROCH
Loser	[(0 1) 0 0 0]	* * *			* * *	*
Winner	[0 0 (0 1) 0]	*		* *	*	*

Table 6
Loser/winner pair before the second demotion

		A-F-R	IAMB	A-F-L	MAIN-R	TROCH
Loser	[0 0 0 (0 1)]			* * *		*
Winner	[0 0 0 (1 0)]		*	* * *		

ing to the corresponding underlying form of 5 syllables, getting [(0 1) 0 0 0]. The mismatch between the production-directed parse and the overt form indicated that the learner's ranking was incorrect. The learner then applied constraint demotion, demoting ALL-FEET-LEFT.

Now that constraint demotion has changed the ranking, the same overt form can be re-interpreted, using the new ranking. As shown in Table 6, the interpretation provided by interpretive parsing is now [0 0 0 (1 0)], and the optimal description for the underlying form of five syllables is [0 0 0 (0 1)]. Again, there is a mismatch, indicating that the learner's ranking is incorrect. Observe, however, that the best interpretation of the overt form now matches the structural description of the target grammar. This occurred despite the fact that the interpretation of the previous step was wrong; the algorithm used an incorrect target but nevertheless moved in the right direction.

Table 7
Loser/winner pair after the second demotion

		A-F-R	A-F-L	MAIN-R	TROCH	IAMB
Loser	[0 0 0 (0 1)]		* * *		*	
Winner	[0 0 0 (1 0)]		* * *			*

Because of the mismatch, constraint demotion is now applied again, using the interpretive parse as the winner and the production-directed parse as the loser. The violations for the new winner and loser are shown in Table 6. The highest-ranked constraint violated more by the loser is TROCHAIC. The constraint violated more by the winner, IAMBIC, is demoted to the stratum immediately below the one occupied by TROCHAIC, the result being as depicted in Table 7. With the resulting ranking, both interpretive and production-oriented parsing give the same (correct) structural description. The learner has succeeded in learning a ranking generating this overt form, and in fact the target language.

The learning procedure is given in Table 8. At any given time, there is a hypothesis ranking held by the learner. Given an overt form, interpretive parsing is used to determine the optimal interpretation of that overt form (under the learner's ranking). That full structural description includes an underlying form. Production-directed parsing is applied to that underlying form to obtain the structural description assigned to the underlying form by the learner's ranking. If the optimal interpretation

Table 8
The iterative learning procedure

0. Start with a hypothesis ranking H_0 and an overt form F.
1. Apply interpretive parsing, using H_x, to F, getting interpretation D_i.
2. Apply production-directed parsing to the underlying form of D_i, getting D_p.
3. If D_i=D_p, learning is done for form F, and H_x is kept.
4. If $D_i \neq D_p$:
   (a) apply constraint demotion to H_x using D_i and D_p, getting new ranking H_{x+1}.
   (b) repeat the procedure from step 1, using the new ranking H_{x+1}.

of the overt form matches the optimal description of the underlying form, the ranking is not changed. So far as can be determined from this overt form, the learner's ranking is correct. If, on the other hand, the interpretation of the overt form given by interpretive parsing does not match the description of the underlying form generated by production-directed parsing, an error has occurred. The learner presumes the interpretive parse to be the correct analysis of the overt form, and applies constraint demotion, with the interpretive parse as the winner and the production-directed parse as the loser. The learner then adopts the resulting new ranking, and the same procedure may be repeated, with either the same overt form or other overt forms.

What results is an iterative procedure that alternates structure assignment (interpretive parsing) and grammar learning (changing the ranking). A hypothesis ranking is used to estimate the hidden structure for an overt form. This hidden structure is then used to determine a new ranking. The new ranking is then used to determine a new estimate of hidden structure, and so forth. Learning is successful if the iterations converge: the assigned interpretations of the overt forms are all grammatical, indicating that the ranking is consistent with the overt forms.

The intuition behind this strategy is that even when the current hypothesis ranking is wrong, the best interpretation of the overt structure is likely to be informative, because it is constrained to match the observed overt structure; even when the best interpretation is itself incorrect, treating it as correct (at least temporarily) can allow the learner to make progress. This is a variation on an idea used in statistical learning. A general class of algorithms that uses this sort of iterative approach to dealing with hidden structure is the class of expectation-maximization, or EM, algorithms (Dempster et al., 1977). EM deals with missing variable values (analogous to hidden structure) by using a guess at a model (analogous to a grammar) to estimate the missing values. The data, including the estimates for the missing values, are then used to select a new model, and the procedure is iterated. The procedure here proposed for Optimality Theory differs from EM in that it is non-statistical, but shares the higher-level outline of using a model/grammar to estimate values for hidden structure, and then using those estimated values to select a new model/grammar.

4. Experiments with metrical stress

To test the iterative strategy, simulation experiments were run. While the system's 11 constraints admit 11! = 39,916,800 distinct total rankings, many of the rankings

generate identical stress patterns. There are a total of 104 distinct languages, counting only languages generated by at least one total ranking of the constraints. The learning algorithm was tried on the overt forms of each of the 104 languages.

4.1. Failure to converge

It is certainly possible for the algorithm to reach a state from which it cannot converge on a correct hierarchy. An interesting case of this is when interpretive parsing produces a description which not only does not match the correct description, but in fact cannot be optimal under *any* ranking. An example of such a description is [(1 0) (0 2) 0]. Notice that this description has inconsistent footing: it has one iambic and one trochaic foot. It is an empirical fact of the OT system described in section 2 that no optimal description can have inconsistent footing. This is due to the fact that 9 of the 11 constraints govern the position of the feet, while the other two constraints, IAMBIC and TROCHAIC, are the only ones determining foot form. Such a description can be the product of interpretive parsing with a ranking having both IAMBIC and TROCHAIC at or near the bottom.

Consider an example with a ranking including the relationships PARSE-SYLLABLE ≫ ALL-FEET-LEFT ≫ IAMBIC ≫ TROCHAIC, and overt form [1 0 0 2 0]. Interpretive parsing gives [(1 0) (0 2) 0], the inconsistent foot form suffered in order to minimize the violation of ALL-FEET-LEFT}. The correct description (the one assigned by the target grammar) is not selected because it has an extra violation of ALL-FEET-LEFT}. The optimal description of 5 syllables under the current ranking (the loser) is [(0 1) (0 2) 0]. As shown in Table 9, the only difference between the interpretive parse (the winner) and the loser is that the winner has a violation of IAMBIC, while the loser has more violations of TROCHAIC. Constraint demotion attempts to improve matters by demoting IAMBIC to below TROCHAIC. But that simply changes the type of foot occurring in the loser: as shown in Table 10, the optimal description is now [(1 0) (2 0) 0], with both feet trochaic. The winner, determined by interpretive parsing, is unchanged under the new ranking. Thus, the winner now has more violations of TROCHAIC. Constraint demotion then demotes TROCHAIC back to below IAMBIC, and the algorithm is effectively back where it started. The result is a cyclic pathology where the learning algorithm keeps demoting IAMBIC and TROCHAIC below each other, without making progress.

4.2. Performance

Giving a complete specification of an algorithm for this learning strategy requires decisions about a number of details. Each language consisted of six overt forms, one for each word of length 2 through 7 syllables. The algorithm used for these simulations works on one overt form at a time, applying up to 5 iterations of the procedure given in Table 8. If success on a form is not achieved within 5 iterations, the algorithm abandons (temporarily) that form and proceeds to the next, without undoing the demotions of the previous 5 iterations. In this fashion, the algorithm makes a pass through the overt forms, starting with the shortest and proceeding by increasing

Table 9
The winner prompts the learner to demote Iambic

		PARSE	A-F-L	IAMB	TROCH
Loser	[(0 1) (0 2) 0]	*	* *		* *
Winner	[(1 0) (0 2) 0]	*	* *	*	*
Correct	[(1 0) 0 (2 0)]	*	* * *	* *	

Table 10
The winner prompts the learner to demote Trochaic

		PARSE	A-F-L	TROCH	IAMB
Loser	[(1 0) (2 0) 0]	*	* *		* *
Winner	[(1 0) (0 2) 0]	*	* *	*	*
Correct	[(1 0) 0 (2 0)]	*	* * *		* *

length to the longest. It makes up to 5 passes through the set of six overt forms. If it makes a pass through the list without making any demotions, then the language has been successfully learned. If the language is not learned within 5 passes through the forms, the algorithm quits and declares failure on that language.

If interpretive parsing produces several descriptions tied for optimality, the learning algorithm simply picks whichever one is the first on the list returned by the parsing algorithm. This naive strategy is rather simple to implement, but makes the algorithm vulnerable to biases in the order in which the parsing algorithm constructs the list of tied parses. However, such ties proved to be rare. Ties among optimal candidates for production-directed parsing are more common, when many constraints are in the same stratum. The loser is always selected by selecting the first member of the list of descriptions returned by production-directed parsing that is not a member of the list of interpretive descriptions returned by interpretive parsing.

Table 11
Algorithm performance

Starting hierarchy	# Languages	# Converged	Median # demotions	Range
Monostratal	104	87	6	2-9
Designed	104	104	6	2-10

Two particular starting hierarchies were used. One, referred to as the *monostratal* hierarchy, had all 11 constraints sharing a single stratum. The algorithm converged on a correct hierarchy for 87 of the 104 languages, when starting from the monostratal hierarchy (the results are shown in Table 11). After examining the nature of the failures, a different starting hierarchy was designed for the purpose of avoiding the traps causing the failures. That hierarchy, referred to as the *designed* hierarchy,

had the two foot-form constraints, TROCHAIC and IAMBIC, in the top stratum, with all of the other constraints in the second stratum. The designed hierarchy resulted in successful learning for every possible language of the system. The success of this initial hierarchy derives from the fact that the only ambiguity an overt form can exhibit in this system is in the form of the foot supporting an overt stress. Starting with the foot form constraints at the top ensures that the foot form constraint dominant in the language stays at the top of the ranking. Because it dominates all of the other constraints, consistent footing is enforced, even at the expense of violating other constraints. This avoids the type of problem illustrated in section 4.1. Provided that the correct foot form constraint is at the top, every interpretation of an overt form will be correct, and the proven properties of error-driven constraint demotion ensure that a correct ranking will be reached. The designed initial hierarchy deals with the ambiguity without stipulating consistent footing within GEN, or even committing to an initial 'unmarked' foot preference.

The use of a specific initial hierarchy raises other issues. While reliance on a specific initial hierarchy may be appealing when it works, it also has some less desirable properties. Heavy reliance on the effects of an initial hierarchy can make a learning algorithm rather fragile, easily knocked off track by a few erroneous observations. It is also unlikely to be a very general solution; the designed initial hierarchy described above was constructed from observations specific to the particular optimality-theoretic system used, with no guarantee of generalization to other OT grammar systems with different candidates and constraints. A more general formal result on the relationship between the initial hierarchy and learning performance would be desirable.

A significant property exhibited by these simulations is the speed of learning. On the correctly learned cases, the algorithm always converged on a correct ranking after at most 10 instances of demotion, fewer than the number of constraints (the illustration in section 3 would count as two demotion instances). An exhaustive search would require checking 11! = 39,916,800 total rankings. That this simple implementation of the iterative strategy converges so quickly attests to the value of the information contained in the interpretive parse, information available in virtue of the structure of Optimality Theory.

5. Discussion

The work presented in this paper is the latest step in a research program investigating the learnability implications of Optimality Theory. This work focused on the challenge of the underdetermination of complete structural descriptions by overt forms. The proposal is to exploit the ordering imposed on structural descriptions in order to estimate the full structural description of an overt form. This is done by selecting, from among those full descriptions consistent with an observed overt form, the one that is optimal with respect to the learner's ranking. This procedure has the intuitive advantage of being the same as the procedure used by a competent speaker to interpret an overt form. In using interpretive parsing, the learner is trying to understand what it is hearing as part of learning from it. The claim is that this

interpretation provides useful information for learning, even when the overt form is inconsistent with the learner's current grammar.

The approach to learning advocated in this paper is motivated by more than the articulation of a learning story for Optimality Theory. It is motivated by the goal of avoiding a pair of extremes in learnability work. One extreme is random search, including the triggering learning algorithm (Gibson and Wexler, 1994) and genetic algorithms (Clark and Roberts, 1993; Pulleyblank and Turkel, to appear). Under this kind of approach, the linguistic theory specifies the space of possible grammars, but that is the extent of its contribution to learning. When confronted with evidence in conflict with the learner's current grammar, the learner randomly selects an alternative grammar to try. The general, uninformed nature of the search can be reflected in long convergence times. The other extreme is embodied in the cue learning approach (Dresher and Kaye, 1990), which is directly constructed upon all the substantive particulars of a specific analysis of a specific linguistic domain. Cue learning requires the learner to be endowed not just with the central principles of universal grammar, but a number of additional substantive facts and relationships spelling out consequences of the principles of the theory. The learning accounts of different linguistic domains must be spelled out separately, each with specific cue-scanning mechanisms separate from the general use of the grammar (in fact, the learner is required to remain oblivious to the actual quality of their language use during learning). These extremes may be symptomatic of the principles and parameters framework itself, which apparently gives little structure for the learner to work with, beyond a specification of a space of possible grammars.

The iterative learning strategy is entirely general to Optimality Theory; no stress-specific knowledge is included in the strategy itself. The frequent and rapid convergence of the simulations using the monostratal initial hierarchy suggest that the iterative learning strategy can play a central role in an overall account of language learning, even if a complete account ultimately requires enhancement by some additional domain-specific elements (such as a designed initial hierarchy). In the broader context of language learnability research, this means that it is possible for the structure of the linguistic theory to provide direction to the learner's search of the space of possible grammars. The formal structure of Optimality Theory makes it possible to interpret overt forms in a meaningful way. The comparison of the best interpretation of the overt form with the currently optimal description of the corresponding underlying form gives the learner a direction to follow within the space of possible grammars. When an observed overt form conflicts with the learner's current grammar, the learner can do better than to randomly select a different grammar; the learner can use the structure of the theory to point to a good alternative grammar, one that is likely closer to the correct grammar. This work shows that the information made available by the structure of Optimality Theory can make a contribution to language learning that is significant indeed.

References

Chomsky, Noam, 1981. Lectures on Government and Binding. Dordrecht: Foris.

Clark, Robin and Ian Roberts, 1993. A computational model of language learnability and language change. Linguistic Inquiry 24, 299–345.

Daelemans, Walter, Steven Gillis and Gert Durieux, 1994. The acquisition of stress: A data-oriented approach. Computational Linguistics 20(3), 421–451.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the *EM* algorithm. Journal of the Royal Statistical Society B, 39, 1–38.

Dresher, B. Elan, 1996. Charting the learning path: Cues to parameter setting. Ms., University of Toronto, revised version to appear in Linguistic Inquiry.

Dresher, B. Elan and Jonathan Kaye, 1990. A computational learning model for metrical phonology. Cognition 34, 137–195.

Ellison, T. Mark, 1994. Phonological derivation in optimality theory. In: Proceedings of the Fifteenth International Conference on Computational Linguistics, 1007–1013.

Furby, Christine, 1974. Garawa phonology. Papers in Australian Linguistics 7, 1–11.

Gibson, Edward and Ken Wexler, 1994. Triggers. Linguistic Inquiry 25, 407–454.

Gupta, Prahlad and David Touretzky, 1994. Connectionist models and linguistic theory: Investigations of stress systems in language. Cognitive Science 18(1), 1–50.

Hammond, Mike, 1995. Syllable parsing in French and English. Ms., University of Arizona, Tucson (ROA-58).

Hayes, Bruce, 1980. A metrical theory of stress rules. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge. [Revised version published by Garland Press, New York 1985].

Hayes, Bruce, 1995. Metrical stress theory: Principles and case studies. Chicago, IL: The University of Chicago Press.

Liberman, Mark and Alan Prince, 1977. On stress and linguistic rhythm. Linguistic Inquiry 8, 249–336.

McCarthy, John and Alan Prince, 1993. Generalized alignment. In: Geert Booij, Jaap Van Marle (eds.), Yearbook of Morphology, 79–154. Dordrecht: Kluwer.

Prince, Alan and Paul Smolensky, 1993. Optimality Theory: Constraint interaction in generative grammar. Technical report, TR-2, Rutgers University Cognitive Science Center, and CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder.

Pulleyblank, Douglas and William J. Turkel, to appear. The logical problem of language acquisition in optimality theory. In: Proceedings of the Workshop on Optimality Theory in Syntax: Is the Best Good Enough? Cambridge, MA: MIT Press.

Smolensky, Paul, 1996. On the comprehension/production dilemma in child language. Linguistic Inquiry 27, 720–731 (ROA-118).

Tesar, Bruce, 1995. Computational Optimality Theory. Ph.D. thesis, University of Colorado, Boulder, June. (ROA-90).

Tesar, Bruce, 1996. Computing optimal descriptions for optimality theory grammars with context-free position structures. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, 101–107.

Tesar, Bruce, 1998. Learnability in Optimality Theory. Linguistic Inquiry 29(2).

Tesar, Bruce and Paul Smolensky, 1998. Learnability in Optimality Theory (long version). Technical report JHU-CogSci-96-4, Department of Cognitive Science, The Johns Hopkins University. (ROA-156).