# Charting the Learning Path: Cues to Parameter Setting

*Bezalel Elan Dresher*

This article argues for an approach to grammar acquisition that builds on the cue-based parametric model of Dresher and Kaye (1990). On this view, acquisition proceeds by means of an ordered path, in which cues to parameters become progressively more abstract and grammar-internal. A learner does not attempt to match target forms (contra Gibson and Wexler 1994), but uses them as evidence for parameter setting. Cues are local, and there is no global fitness metric (contra Clark and Roberts 1993). Acquisition of representations and acquisition of grammar proceed together and cannot be decoupled in the manner of Tesar and Smolensky (1998).

*Keywords:* learnability, parameters, language acquisition, metrical phonology

Current approaches to the problem of learnability of grammars assume a highly constrained theory of Universal Grammar (UG), within which crosslinguistic variation is kept to certain limits. These limits are set, depending on one's theory, either by a series of variable parameters that learners must fix at their correct values (Chomsky 1981) or by a series of constraints that learners must correctly rank (Prince and Smolensky 1993). An explanatory theory ought to specify how the learner sets the parameters or ranks the constraints on the basis of relevant input data.

Two fundamental problems must be overcome in developing a learning model. The first is that parameters and constraints interact in complex ways, and it is difficult to reliably discern what specific contribution each one makes to the whole. A learner whose hypothesized grammar does not successfully account for the target input would have no reliable information about the nature of the error. We can call this the *Credit Problem*.[1] A second fundamental problem is that parameters and constraints are stated in terms of abstract entities that the learner is not initially able to identify. For example, metrical theory is couched in terms of concepts such as heavy syllable, head, constituent, and projection. These entities do not come labeled as such in the input,

I would like to thank Janet Fodor, Ted Gibson, Norbert Hornstein, Alana Johns, David Lightfoot, Alan Prince, Bruce Tesar, Ken Wexler, an anonymous reviewer, and audiences at MIT, the University of Toronto, SUNY Buffalo, WECOL 1994 at UCLA, the Maryland Mayfest 95 at the University of Maryland at College Park, and CTIP 95 at the Abbaye de Royaumont for useful comments on earlier versions of this article. Needless to say, none of the above necessarily agrees with the views expressed here. I am grateful for the support of SSHRC research grants 410-92-0885 and 410-96-0842, and I would like to thank my co-investigator in these projects, Keren Rice, and the students and colleagues who have collaborated with us, for their suggestions and encouragement.

[1] Clark (1989) calls this the *Selection Problem*.

Linguistic Inquiry, Volume 30, Number 1, Winter 1999
27–67

but must themselves be constructed by the learner. Since parameters are stated in terms of metrical theory, whereas the cues to these parameters must be stated in terms of observable data, what the correct cue to a given parameter is must be empirically determined (the same holds if the problem is construed as one of constraint ranking). We can call this the *Epistemological Problem*.

(1) *Two fundamental problems*
   a. *The Credit Problem:* When there is a mismatch between a target form and a learner's grammar, there is no way of reliably knowing which parameters/constraints must be reset to yield a correct output.
   b. *The Epistemological Problem:* There is a gap between the vocabulary in terms of which parameters/constraints are couched and the learner's analysis of the input.

These problems make it a challenge to devise a reliable procedure that guarantees that the learner will converge on the target grammar.

The cue-based learning model (called YOUPIE) of Dresher and Kaye (1990), which is a learning model for a parametric version of metrical phonology, was designed as an attempt to overcome these problems in one area of phonology, though the principles are intended to hold in other domains also. In section 1, I sketch some general properties of the model and show how it works in an example case. Further implications of the theory are discussed in section 2. In sections 3 through 5, I consider alternative approaches that have recently been proposed: the Triggering Learning Algorithm (Gibson and Wexler 1994), a genetic algorithm (Clark 1990, 1992, Clark and Roberts 1993), and Robust Interpretive Parsing/Constraint Demotion (Tesar 1997, 1998, Tesar and Smolensky 1996, 1998). I will argue that each of these learning models fails to adequately address one or both of the fundamental problems. Though some of these models deal with phonology and others with syntax, the fundamental problems discussed in this article remain the same, and I will assume that, for purposes of this discussion, what holds in one domain holds in the other as well. Section 6 is a brief conclusion.

## 1 A Cue-Based Learner (Dresher and Kaye 1990)

Some of the main features of the Dresher and Kaye 1990 model are listed in (2).[2]

(2) *Properties of a cue-based learner (Dresher and Kaye 1990)*
   a. UG associates every parameter with a *cue*.
   b. A cue is not an input sentence or form but is something that can be derived from input.
   c. Cues must be *appropriate* to their parameters in the sense that the cue must reflect a fundamental property of the parameter, rather than being fortuitously related to it.
   d. What the correct cue to any given parameter is must be empirically determined (by the linguist—not the learner, to whom it is supplied by UG). There is thus no parameter-independent general algorithm for parameter setting.

[2] For further discussion of various aspects of this learning model, see also Dresher 1992, 1993, 1994.

e. Parameter setting proceeds in a (partial) order set by UG: this ordering reflects dependencies among cues and specifies a learning path (Lightfoot 1989). The setting of a parameter later on the learning path depends on the results of setting earlier ones.
f. A parameter that has a default state remains in it until the learner detects its cue, which acts as the trigger to move to the marked setting. Symmetrical parameters (e.g., directional parameters) may have positive cues for both values.
g. The learning strategy is loosely speaking ''deterministic,'' in the sense of Marcus (1980) and Berwick (1985), in that the learner may not backtrack or undo parameter settings that have already been set. Some such restriction is necessary if the learner is to be prevented from getting into infinite loops.[3]
h. Determinism does not hold in the following case: when a parameter is set to a new value, all parameters that depend upon it (follow it in the order) revert to default.
i. Cues are local in the sense that each decision depends on finding a specific configuration in the input, which the learner acts on without regard to the final result. Hence, learners are not trying to match the input.[4]
j. Cues become increasingly abstract and grammar-internal the further along the learning path they are.

By way of illustration, consider the core stress system of English, which for purposes of this example we can consider to be the same as Latin. This stress pattern can be characterized as in (3); some words illustrating this pattern are shown in (4).

(3) *English/Latin stress*
Main stress falls on the penultimate syllable if it has a long vowel or is closed by a consonant; otherwise, main stress falls on the antepenultimate syllable.

(4) *Examples*
a. álgebra, Cánada, génesis, América
b. Vàncóu:ver, aró:ma, horí:zon, Mànitó:ba
c. agénda, appéndix, Hèlsínki, Bònavénture

Following standard accounts (e.g., Halle and Vergnaud 1987), the metrical patterns of sample words are derived from grid representations such as those in (5).

(5) *Acquired representations*

| a. | b. | c. | |
|---|---|---|---|
| x | x | x | Line 2 |
| (x) | (x x) | (x) | Line 1 |
| x (x x)⟨x⟩ | (x x)(x)⟨x⟩ | x(x) ⟨x⟩ | Line 0 |
| L L L L | L L H L | L H L | Syllables |
| Ameri ca | Mani to:ba | agenda | |

[3] See Nyberg 1991a,b for detailed discussion of the merits and drawbacks of determinism. He argues for a limited nondeterministic learning model.

[4] It is important to underscore that I do not mean ''local'' in the sense of being confined to a particular vicinity that is visible at the surface. The intended sense should become clear in the light of the examples that follow.

In these grids *H* represents a heavy syllable (a syllable containing a long vowel or one that is closed by a consonant), and *L* represents a light syllable (a syllable containing a short vowel). The relative stress of a syllable is indicated by the height of its grid column. Parentheses indicate constituent boundaries. Angle brackets indicate an extrametrical syllable, that is, a syllable that does not participate in the computations of the metrical grid (''outside the meter''). In each line 0 constituent, one and only one element projects a mark on line 1: this element is the head of the line 0 constituent. Line 1 marks are similarly gathered into a constituent whose head is on line 2.

Let us assume that the grids in (5), constructed in accordance with parameters that we will take up as we proceed, are what learners of English have to arrive at. Thus, on the theory assumed here, the rule in (3) is a descriptive generalization without status in the grammar to be attained; rather, the pattern described by (3) will be seen to follow from the appropriate setting of the parameters. I assume also that the input that the learners have to work with consists of words associated with primitive grids that represent only the observed stress contours of each word. For the words in (5), the input (i.e., the learner's representation of the surface form) would look like (6).

(6) *Initial representations*

| a. | | b. | | c. | | |
|---|---|---|---|---|---|---|
| x | | x | | x | | Line 2 |
| x | | x x | | x | | Line 1 |
| x x x x | | x x x x | | x x x | | Line 0 |
| S S S S | | S S S S | | S S S | | Syllables |
| America | | Manito:ba | | agenda | | |

The input grids indicate the shape of the stress contour of a word, but they lack constituent boundaries and extrametricality markings; these must be supplied by the learners. Also, since the distinction between heavy and light syllables is not self-evident to begin with, *L* and *H* are replaced by *S*, which represents any syllable.[5]

### *1.1 Quantity Sensitivity*

In English the location of stress depends on the distribution of heavy syllables, as well as location in the word. Hence, a learner can make no progress in acquiring the correct pattern without first determining that English distinguishes light from heavy syllables; that is, English stress is quantity sensitive (henceforth QS). Stress systems that do not distinguish between syllable types are called

[5] The formulation of this learning problem incorporates a number of assumptions about the state of the learner (see Dresher and Kaye 1990 for further discussion). I am assuming that learners are able to treat the words in question as distinct units, that they have learned to map the appropriate acoustic cues into primary and secondary stress, and that they have already acquired enough syllable structure to parse words into syllables correctly enough for purposes of stress. Also, I assume for purposes of this example that the lexical (underlying) segmental representations of the words to be stressed do not differ significantly from their surface forms. The problem of acquiring underlying representations requires information from the morphology, among other things; I assume it involves rather complex interactions among various components of the grammar.

**Table 1**
Word classes in quantity-insensitive (QI) and quantity-sensitive (QS) systems

| | QI: Syllable = S | QS: Syllable = H or L |
|---|---|---|
| 2-syllable words | {S S} | {L L} {H L} {L H} {H H} |
| 3-syllable words | {S S S} | {L L L} {H L L} {L H L} {H H L} {L L H} {H L H} {L H H} {H H H} |
| 4-syllable words | {S S S S} | {L L L L} {H L L L} {L H L L} . . . |

quantity insensitive (QI). The task, then, is to discover that English stress is QS without making use of the generalization in (3), since this pattern cannot itself be discerned until one distinguishes between light and heavy syllables.

One operation that is available to a learner at this early stage in acquiring the system is classification. It is reasonable to suppose that learners begin with simple representations and must be driven to adopt more complex ones. Thus, we may suppose that the default is to assume that all syllables are the same for purposes of stress, that is, to assume that stress is QI. Because all syllables have the same status in QI systems, it follows that words with the same number of syllables are all alike from the point of view of the metrical parameters. In QS systems, by contrast, this is not the case, as is demonstrated by the equivalence classes of word types shown in table 1.

In QI systems all words with *n* syllables should have the same stress contour, since they are all effectively equivalent. Taking quantity insensitivity to be the default case, a learner will continue to assume that stress is QI until it encounters evidence that words of equal length can have different stress contours.

(7) *Quantity (in)sensitivity*
   a. *Parameter:* The language {does not/does} distinguish between light and heavy syllables (a heavy syllable may not be a dependent in a foot).
   b. *Default:* Assume all syllables have the same status (QI).
   c. *Cue:* Words of *n* syllables, conflicting stress contours (QS).

Such evidence is abundant in English, as is apparent in (4). For example, the trisyllabic words in (4a) have initial stress, conflicting with the trisyllabic words in (4b) and (4c), which have stress on the middle syllable; similarly, *América* conflicts with *Mànitóba,* and so on. The existence of conflicting stress contours on a wide scale would lead the learner to abandon the default hypothesis. Note that quantity sensitivity is not the only cause of such conflicts: the language in question may have lexical accent, for example. A fuller specification of the learning path would have to include means for distinguishing between quantity sensitivity and lexical accent, but I cannot consider all the possibilities here (see Dresher 1994 for some discussion). Similar considerations hold all along the line. Assuming, though, that other possibilities are ruled out, the learner is led to revise the input representations, now distinguishing between light and heavy syllables.

Here, too, there are choices to make, because not every language has the same characterization of what a heavy syllable is. Some languages do not count closed syllables with short vowels as heavy. (8) gives a slightly oversimplified picture of the possibilities, but one I will adopt here: I will assume that syllables that end with a short vowel (short open syllables) are universally light, and that syllables with long vowels are universally heavy. Closed syllables may go either way (here and elsewhere, a period indicates a syllable boundary).

(8) *Light and heavy syllables*

| *Always light* | *Light or heavy* | *Always heavy* |
|---|---|---|
| . . . V. | . . . VC. | . . . VV |

In order to determine which style of quantity sensitivity English adopts, we can continue with the classification test used to diagnose quantity sensitivity in the first place. Let us assume that when a learner determines that a language is QS, it revises its initial representations, now characterizing syllables as being either light or heavy. Suppose that the initial revision incorrectly assumes that closed syllables are light; the learner would arrive at the word classes in (9).

(9) *Assuming QS stress, closed syllables light: conflicting words*[6]

a. L L L ál.ge.bra, a.gén.da, Hèl.sín.ki

b. L L L L A.mé.ri.ca, Bò.na.vén.ture

The new representations still contain conflicting words: thus, words of the pattern *L L L* do not all have the same stress contour, nor do words of the pattern *L L L L.* These conflicts, which would again exist on a large scale in the language, would serve as a trigger for the learner to try the other possibility in (8), which leads to representations in which closed syllables count as heavy (10); these representations contain no conflicts, an indication that they can serve as a basis for proceeding to set further metrical parameters.

(10) *Assuming QS stress, closed syllables heavy: no conflicting words*

| | | | |
|---|---|---|---|
| H H H | Vàn.cóu:.ver | H H L | Hèl.sín.ki |
| H L L | ál.ge.bra | L L H | gé.ne.sis |
| L L L | Cá.na.da | L H L | a.ró:.ma, a.gén.da |
| L H H | ho.rí:.zon, a.ppén.dix | L L L L | A.mé.ri.ca |
| L L H L | Mà.ni.tó:.ba | L L H H | Bò.na.vén.ture |

Having found the heavy syllables, what the learner knows about the sample words in (6) is given in (11).

[6] Periods mark syllable boundaries. Note that the final vowel in *Helsinki* is long (or tense) in most dialects of English (but not in the English of the southern United States). I assume that the vowel is underlyingly short and remains so for purposes of stress assignment, and that it is lengthened by a rule requiring that all final nonlow vowels be long at the surface. A learner who does not know this will designate *Helsinki* at this point as having the syllable pattern *L L H*.

(11) *New representations with light and heavy syllables*

| a. | b. | c. | |
|---|---|---|---|
| x | x | x | Line 2 |
| x | x x | x | Line 1 |
| x x x x | x xx x | x x x | Line 0 |
| L LLL | L LH L | LH L | Syllables |
| America | Manito:ba | agenda | |

Of course, determining that a stress system is QS means more than just assigning *H*s and *L*s to syllables. Heavy syllables have a particular status with respect to how they are represented on the metrical grid: the basic intuition is that heavy syllables are inherently more prominent than light syllables. There are a number of ideas about how this prominence is expressed. For purposes of this discussion, I will assume that UG requires that a heavy syllable project a mark on line 1 of the grid (Prince 1983), as shown in (12).[7]

(12) *Heavy syllables project a line 1 grid mark*

| a. | b. | c. | |
|---|---|---|---|
| x x | x | x x x | Line 1 |
| x x x | x x x x | x x x | Line 0 |
| LH H | L LH L | H H H | Syllables |
| hori:zon | Manito:ba | Vancou:ver | |

The grids in (12) represent constructions of the learner who has set the parameter for quantity sensitivity, but not the other parameters of metrical theory. We observe that these representations are quite remote from the target surface contours: quantity sensitivity accounts only for prominence of heavy syllables, but the word *Manitoba* contains a light syllable that has a secondary stress. This additional stress must be assigned by other parameters. Conversely, not every heavy syllable is stressed equally. For example, in *Vancouver* there are three heavy syllables: the first has a secondary stress, the second has primary stress, and the third has no stress at all. The difference between primary and secondary stress is expressed on line 2; abstracting away from this difference, the line 1 marks of these heavy syllables are consistent with the observed stress contours, since a line 1 mark translates into some kind of stress. The stressless final heavy syllable, though, is contrary to what we might expect given the representation in (12), since its line 1 mark should give it some stress. Recall that final syllables in (5) are marked as being extrametrical, which accounts for their lack of stress. The discrepancy between the forms in (12) and the target contours will eventually be resolved when other relevant parameters are set. The learner must not panic at this point, even though it is positing a stress that is not consistent with the target. There are

[7] Idsardi (1992) and Halle and Idsardi (1995) propose instead that heavy syllables are associated with a parenthesis on line 0, meaning that a heavy syllable must begin or end a line 0 constituent. This formalism entails a slight change to the learning algorithm (see Dresher 1994).

such discrepancies everywhere: in the word *genesis,* for example, the only heavy syllable, being final, receives no stress at all. In fact, in a word like *genesis* the adoption of quantity sensitivity might seem to be a step in the wrong direction. However, by our assumption, the learner is not concerned with how well the result of parameter setting leads to matches with target forms. Since the cue for quantity sensitivity indicates that English is QS, this is what the learner concludes, regardless of the outcome in any particular form.

### *1.2 Main Stress*

Let us move on to consider the location of main stress. Main stress is assigned by promoting either the leftmost or rightmost line 1 mark onto line 2. So, although main stress is not confined to the first or last syllable, it is limited to the first or last line 1 mark, which is the head of the first or last line 0 constituent.

(13) *Main stress*
   a. *Parameter:* Project the {left/right}most element of the line 1 constituent.
   b. *Cue:* Scan a constituent-sized window at the edge of a word. Main stress should consistently appear in either the left or right window.

Looking at some English words, we might think that it is a simple matter to figure out how this parameter is set in English. For example, if we look at a target contour like that of *Manitoba* in (11b), we observe that the rightmost line 1 mark is more prominent, so main stress must be on the right. This is the right answer, but a learner cannot rely on target contours in every language. Some languages have no discernible secondary stress, meaning that every word has only one surface stress. In such languages, surface contours of individual words taken in isolation are of no help to the learner.

An example of such a language is Selkup (Halle and Clements 1983). Its stress pattern is described informally in (14).

(14) *Selkup stress*
   Stress falls on the rightmost long vowel; if there is no long vowel, stress falls on the initial syllable.

Some sample words are given in (15).

(15) *Selkup examples*

| | | | |
|---|---|---|---|
| qúmmin | ‘human being (gen.)’ | kanaˆmíː | ‘our dog’ |
| qummɪː | ‘our friend’ | qólʸcimpati | ‘found’ |
| ámirna | ‘eats’ | uːcikkóːqɪ | ‘they two are working’ |
| qumóːqɪ | ‘two human beings’ | qumoːqlɪlɪː | ‘your two friends’ |

Selkup has a QS stress system where only long vowels count as heavy. Since there is only one stress per word, inspection of surface contours gives no indication about which side main stress is on, as we see in (16a).

(16) a. *Selkup surface contour*

| | |
|---|---|
| x | |
| x | |
| x x x x | |
| H L H L | |
| u:cikko:qɪ | |

b. *Selkup hypothesized contour*

| | |
|---|---|
| x | Line 2 |
| x x | Line 1 |
| x x x x | Line 0 |
| H L H L | Syllables |
| u:cikko:qɪ | |

If the learner takes into account the fact that every *H* ideally projects a line 1 grid mark, as in (16b), it can notice that the rightmost line 1 mark is the one promoted to line 2; that is, main stress is on the right in Selkup, again a correct conclusion. However, the same technique applied to English gives misleading or confusing results. For example, in *horizon* the leftmost heavy syllable is promoted, suggesting, incorrectly, that main stress is on the left; in *Vancouver* the middle heavy syllable is promoted, not a theoretical possibility for the main stress parameter. Again, we must take account of extrametricality.

What we conclude from this discussion is that a learner can be misled if it tries to determine the position of main stress either from target surface contours or from hypothesized contours based on setting only the parameter for quantity sensitivity. The problem in both cases is that the representations are too crude and uninformative to give a consistently reliable result. The learner needs more information, particularly about extrametricality and constituency. In the version of metrical theory being assumed here, the first or last line 1 mark is, in most cases, the head of the first or last line 0 constituent. This fact suggests a cue for main stress, given in (13b): scan a constituent-sized window at the edge of a word; main stress should consistently appear in either the left or right window.

### *1.3 Bounded Constituents*

It follows from (13) that the learner does not need to know exactly what the constituents of a word are, or how they come to be there, in order to determine whether main stress is on the left or the right, but it does need to know how big a metrical constituent is. In particular, it needs to know if line 0 constituents are bounded or not; for purposes of this discussion, let us limit bounded constituents to binary ones.

(17) *Bounded constituent construction*
- a. *Parameter:* Line 0 constituents are {unbounded/bounded}.
- b. *Default:* Assume line 0 constituents are unbounded.
- c. *Cue:* The presence of a stressed nonedge light syllable indicates bounded constituents.

If a language has bounded constituents, then a constituent-sized window will not be more than two syllables long. A typical example is Maranungku (Tryon 1970, Hayes 1981).

(18) *Maranungku stress*
Main stress falls on the initial syllable; secondary stress falls on every second syllable thereafter.

This stress pattern can be obtained by building left-headed binary feet, starting from the left, as shown in (19).

(19) *Maranungku: binary constituents starting from the left*

| a. | | b. | |
|---|---|---|---|
| x | | x | Line 2 |
| (x x) | | (x x) | Line 1 |
| (x x)(x x)(x) | | (x x)(x x)(x x) | Line 0 |
| langka rate ti | | wele pene manta | |

By contrast, if a language does not utilize bounded constituents, the only constituents it will have, if it has any, are those created by heavy syllables and by edge rules. The effect of heavy syllables we have already seen; the effect of edges is illustrated in the statement of the Selkup stress rule (14): when there is no long vowel, stress defaults to the initial syllable. Another example is Koya (Tyler 1969, Hayes 1981).

(20) *Koya stress*
Stress falls on the head of every closed or long syllable as well as on the head of the initial syllable. Main stress falls on the initial syllable.

To achieve this stress pattern, left-headed constituents can be constructed on line 0. Each constituent extends until it hits a heavy syllable, which must begin a new constituent, or the end of the word. In this language a word-initial light syllable is the only light syllable that can be stressed.

(21) *Schematic Koya words*

| a. | | b. | |
|---|---|---|---|
| x | | x | Line 2 |
| (x x) | | (x x x) | Line 1 |
| (x x x) (x x) | | (x x) (x x) (x) | Line 0 |
| L L L H L | | L L H L H | Syllables |

English, like Maranungku, has bounded constituents; how might a learner determine this? A number of possible cues come to mind. One is the presence of alternating stress, but this turns out to be a slippery cue (Dresher and Kaye 1990, Dresher 1994:77–78). The essential difference between languages with bounded constituents and languages without them is that in the latter, constituent edges must be associated either with heavy syllables or with the edge of a word. Therefore, the only light syllable that can be stressed is one that is at a word edge, as we have seen in the cases of Selkup and Koya. It follows that the presence of a stressed light syllable that is not at a word edge is evidence for bounded constituents. I adopt this as the correct cue for boundedness, given in (17c).

English has such internal stressed light syllables; an example—actually, the only example in our data set—is the word *América.* Without this word, the forms in (4) would be equally analyzable as belonging to an unbounded stress system with the pattern: stress the last heavy syllable that does not occur in the final syllable; otherwise, stress the initial syllable.

### *1.4 Extrametricality*

The cue for boundedness refers to word edges, and it should now be said that the position of the effective edge of a word is also subject to parametric variation. This is because a peripheral syllable may or may not be designated as extrametrical (Liberman and Prince 1977). This device is available only at an edge of a stress domain.

(22) *Extrametricality*[8]
  a. *Parameter:* A syllable on the {right/left} {is not/is} extrametrical.
  b. *Cue:* Stress on a peripheral syllable rules out extrametricality on that side.

I will assume here that when a syllable is extrametrical, it may project no grid marks above line 0. In English nouns the rightmost syllable is extrametrical, which explains why final heavy syllables have no stress. One might take the constant absence of stress to be the cue for extrametricality, but this does not work in general. For one thing, heavy syllables can be destressed for a variety of reasons, not just because of extrametricality. For another, an edge syllable may be constantly stressless without being extrametrical: for example, if iambic (i.e., right-headed binary) feet are constructed from the beginning of the word, the first syllable will never have stress, but it is not extrametrical either. In general, the absence of stress where it is otherwise expected is rarely a reliable cue; however, the presence of stress where it is unexpected is an important cue. Therefore, the discovery that an edge syllable receives stress is good evidence that extrametricality is excluded on that side, and this is the cue given in (22). In English, for example, we can exclude extrametricality on the left side, because many words have a stress on the initial syllable.

What about the right side? There are no stressed final syllables in (4), which means that extrametricality cannot be excluded; but, as just noted, the constant lack of a stress is not a sufficient cue for extrametricality. In this situation the learner cannot make a final determination until later. In the meantime it must keep open the possibility that there may be extrametricality on the right.

### *1.5 Headedness and Directionality of Feet*

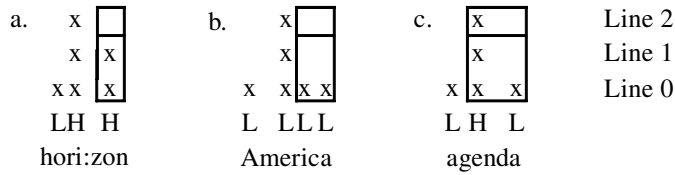
The learner is now ready to determine the setting of main stress. The cue, as we have seen, is to scan a constituent-sized window at each edge of the word, with the expectation that main stress will consistently show up in one of these windows. In English such a window is at most two syllables long; when two heavy syllables are adjacent at an edge, it is only one syllable long. As shown in (23), if no extrametricality is posited, main stress does not appear in the right-hand

[8] This parameter contains two separate choices that are not entirely independent. Assuming that a language may not choose to make both the leftmost and rightmost syllables extrametrical, the possible settings are these: a syllable on the left is extrametrical; a syllable on the right is extrametrical; no syllable is extrametrical.

It has been observed (Hayes 1995:57–58) that extrametricality on the right side is much more common than extrametricality on the left. Prince and Smolensky (1993:51) propose that extrametricality be replaced by NONFINALITY, a (violable) constraint requiring that stress not fall on the final syllable of a word.

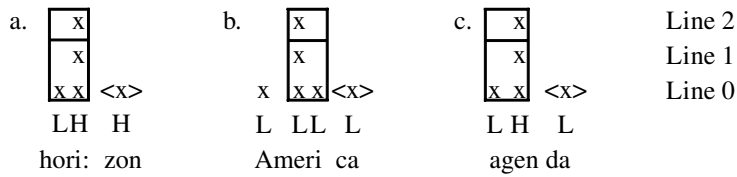
window of many words, like *horizon* and *America*; the only success is *agenda,* if one assumes left-headed feet.

(23) *Constituent-sized window at the right edge, no extrametricality*




However, if extrametricality is assumed, main stress appears consistently in the right-hand window.

(24) *Constituent-sized window at the right edge, with extrametricality*




In this fashion, the learner sets the values of main stress and extrametricality on the right at the same time.

The constituent-sized windows in (23) and (24) have been constructed without knowing exactly what the constituents are, and they are inaccurate in some respects. I am assuming here that English feet are QS trochees, which is to say that they are left-headed maximally binary feet, as shown in (25).

(25) *QS trochees*

a. x . (L L) b. x . (H L) c. x (H) d. x . (L) or (L)

A trochee may consist of two light syllables (25a), or a heavy syllable followed by a light syllable (25b), or a heavy syllable by itself (25c). A single light syllable (25d) is stressed when it carries main stress (in English this situation arises only in words with two syllables, like *puddle*), but not otherwise; hence, the initial syllable of *Vàncóuver* is heavy and stressed, but the initial syllable of *agénda* is light and unstressed. In (24a) and (24c) the window includes (L H), which is not a possible foot in English. The reason it does this is that (L H) is a possible bounded foot in the inventory of UG, in the theory assumed here. Notice that in these cases it does not matter whether the light syllable preceding the heavy is part of the final foot or not, since stress in any case falls on the heavy syllable. In these words, then, the result is the same whether one assumes right-headed or left-headed feet. In (24b), though, headedness does matter: main stress appears in the window only if one assumes that feet are left-headed.

In this language, then, determining main stress has the consequence of determining headedness of feet. This is not true in general, however, as there are cases where main stress does

not determine headedness. In the general case, headedness and directionality of feet—whether feet are constructed from the left or right—must be determined later. The cue proposed in Dresher and Kaye 1990 is shown in (26). Another way of stating the cue is as follows: For each setting of direction and headedness, scan across the word. The presence of a stressed syllable in what should be a weak position rules out a setting; the presence of an unstressed syllable in a strong position does not count.

(26) *Headedness and directionality of feet*[9]
   a. *Parameters:* {Left/right}-headed feet are constructed from the {left/right}.
   b. *Cue:* Scanning from the {left/right}, a light syllable {following/preceding} any other syllable must be unstressed.
   c. *Example:* Scanning from the left, if for all (*X L*), *L* is unstressed, then direction = left, headedness = left. If for all (*L X*) *L* is unstressed, then headedness = right.

Consider how this cue applies to the English examples we have been looking at. If we start from the left, we find that feet are not consistently left-headed, because of words like *América* in (24): the first foot we would build would comprise the two light syllables (*A mé*), a right-headed foot. But neither are feet consistently right-headed, as shown by *Mànitóba* (11)—(*Mà ni*)—or *Cánada* (10)—(*Cá na*). We learn that there is no consistent footing if we scan from the left.

From the right, we do get a consistent result. Consider again *América* in (24). Starting from the right edge, which means excluding the extrametrical final syllable, we form a foot with two light syllables (*mé ri*). This foot is of the form (*X L*), where *L* is unstressed, and so is consistent with left-headedness. We continue to scan leftward: since the stressed *L* was already considered—it was *X* in the (*X L*) foot we just formed—we are left only with the first *L*. Since it does not follow any syllable, no test applies to it. Applying a moving window of the form (*X L*) from right to left in every word in our set, we will see that all feet are left-headed. Hence, English has left-headed feet constructed from the right.

### *1.6 Destressed Feet*

One final source of crosslinguistic variation we will consider here concerns the treatment of feet consisting of single syllables. We have seen that a foot consisting of a single heavy syllable is stressed in English, whereas a foot consisting of a single light syllable is not.[10] There are other possibilities, though: feet may be destressed (or not formed in the first place) in a wide variety

[9] Two independent choices are involved here, one for headedness and another for directionality. As with extrametricality, the choice of left and right in these parameters may not be entirely symmetric. Hayes (1987) has observed that trochees (left-headed binary feet) appear to be associated with quantity insensitivity, whereas iambs (right-headed feet) appear to be associated with quantity sensitivity. On these grounds, he proposes a revised set of asymmetric parameters that build in these dependencies (see also Prince 1990 for further development of this idea, and Kager 1993 for arguments that these asymmetries are derivable from independent rhythmic considerations and should not be built into the parameters themselves).

[10] Alternatively, such syllables may not be parsed into feet at all. The distinction between destressed feet and unparsed syllables is not crucial to this discussion.

**Table 2**
Parameter settings of selected languages

| | English/Latin | Selkup | Maranungku | Koya |
|---|---|---|---|---|
| Syllable quantity | QS | QS | QI | QS |
| Closed syllables | heavy | light | n/a | heavy |
| Main stress | right | right | left | left |
| Bounded feet | yes | no | yes | no |
| Extrametricality | right | no | no | no |
| Foot head | left | left | left | left |
| Foot directionality | right | n/a | left | n/a |
| Destressing | (L) | all but head foot | no | no |

of conditions, which typically are sensitive to syllable weight and the presence of stress clashes (adjacent stresses). The Dresher and Kaye 1990 learner treats these phenomena in a separate destressing module that consists of seven parameters and applies after the fixing of the other parameters discussed above. Without going into the particulars, I show in (27) the general strategy employed by the learner in all these situations. Note that (27) conflates a number of parameters that all have their own separate cues.

(27) *Destressing*
  a. *Parameters:* {Various types of} feet are destressed in {various situations}.
  b. *Default:* All feet are stressed.
  c. *Main cue:* The absence of stress on a foot.
  d. *Example:* The lack of stress on the first syllable of *agénda,* with acquired foot structure (*à*)(*gén*)⟨*da*⟩, shows that this foot is destressed (further cues reveal the conditions under which this occurs).

### *1.7 Summary of Parameters and Sample Languages*

The languages mentioned above have the parameter settings shown in table 2.

### *1.8 Order of Parameter Setting*

We have not considered all the parameters that have been proposed as part of the theory of metrical phonology; but continuing in this fashion, we can go on to specify the entire learning path for acquiring the metrical system of this language.[11] The way this learning model addresses

[11] For recent surveys of stress systems, see Hayes 1995 and Goedemans, van der Hulst, and Visch 1996. For a review of some machine learning techniques applied to the acquisition of stress, see Gillis and Durieux 1996.

Because the cues and ordering of a cue-based learner must be determined empirically, no formal proof of the correctness of the learning model can be offered. Gillis, Durieux, and Daelemans (1995) report on an empirical test of YOUPIE that they conducted, using artificially generated languages consisting of all possible types of words of two, three, and four syllables. They found that it is successful on 75–80% of these languages, depending on how they are counted. Their analysis of the errors shows that many of them relate to problems with extrametricality and foot size. It

the Credit Problem and the Epistemological Problem should by now be clear. The Credit Problem is solved for the learner by associating each parameter with a cue: the learner always knows what to look for to set a parameter. Moreover, the learner is never asked to apportion credit for an entire form to a set of parameters. The Epistemological Problem is solved by ordering the parameters; the parameters discussed above are ordered as in (28).

(28) *Order in which parameters must be set*
   a. *Syllable quantity:* Establishes whether feet are QI (default) or QS (and the type of quantity sensitivity).
   b. *Extrametricality:* Establishes effective edge of domain; can only exclude extrametricality at this point.
   c. *Foot size:* If QI, only bounded feet available; if QS, unbounded is default.
   d. *Main stress:* Depends on correct settings of all the above.
   e. *Headedness:* Sometimes depends on having set main stress.
   f. *Directionality:* Cannot be determined apart from headedness.
   g. *Destressing:* Determined by comparing stresses predicted by above parameter settings with actual stresses.

This ordering allows for a general progression, both in the representations and in the cues, from relatively simple to more complex and more abstract. The cue for quantity sensitivity, for example, coming near the beginning of this learning path, is couched in terms that presuppose little knowledge of any details of the grammar. The learner needs only to be able to keep track of stress contours and syllables. By contrast, the cue for main stress is considerably more sophisticated in what it assumes about the grammar, and the cues for destressing can refer to all aspects of metrical structure. If parameters were unordered, then the cues could not be stated in this progressive fashion.

## 2 Some Issues Pertaining to Ordering

The picture suggested so far is of a learner moving along a path, setting first one parameter, then another. Suppose that the learner at some point is at some parameter—say, number 5. What is the status of parameters further along? Are they all unset? Or do they remain at default? A number of possibilities are compatible with the model sketched above, and I will briefly consider some of them in this section.

### *2.1 Ordering in an Incremental Learner*

Imagine first a learner that collects data for some preset amount of time, or until it decides it has seen everything important, before attempting to set any parameters. Such a model, which has

is possible that the cues for these parameters need to be refined or replaced. Another possibility, however, is that the fault lies in restricting the sample languages to words of fewer than five syllables; it is possible that words of more than four syllables are necessary for the algorithm to learn certain parameter settings. In the absence of further details about which languages were not successfully learned, it is hard to draw any more conclusions from these results.

access to all the relevant data, we can call a *batch* learner (29a). Now contrast this with another possible model, one that operates in *incremental* mode (29b). An incremental learner processes data as they come in, trying to extract as much information as it can from each new datum.

(29) *Two types of learners*
  a. *Batch learner:* Collect all data, then set parameters.
  b. *Incremental learner:* Adjust parameter settings as each datum comes in.

An incremental model appears to lend itself better to a developmental interpretation, since we think of language acquisition in real time as being incremental. However, the situation is not as straightforward as it seems.

A batch processor can simply set the parameters in the indicated order, and it will not go wrong, because it has all the relevant data before it. Therefore, by the time it has to set the parameter for, say, main stress, it will have already correctly set the values for syllable quantity, extrametricality, foot size, and so on. Such a learner is quite powerful: knowing that it has seen all the relevant data gives it a considerable advantage over an incremental learner. But for that reason it also appears less realistic, so let us look at how an incremental learner would deal with these dependencies.

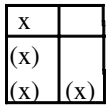
Because an incremental learner is setting parameters on the fly, perhaps before it has encountered all the relevant data, it is important that it not make false moves from which it may not be able to recover. Recall that the cue-based learner keeps the default setting of a parameter until it sees positive evidence to change it to the marked setting. Though this appears to be a prudent strategy, it cannot keep the learner from making false moves in the course of acquiring a set of interacting parameters.

Thus, suppose that the learner is trying to learn Selkup (14)–(16). Recall that Selkup stress is actually QS, where only a long vowel counts as heavy; it has unbounded left-headed feet, and main stress on the right. Suppose the learner has figured out that the language is QS, but let us assume now (contrary to what we did before) that the default setting for a QS language is that closed syllables with short vowels count as heavy. Suppose that the learner has not yet seen any evidence to move from this default state and so is treating closed syllables as heavy in addition to long vowels. Now, there is no problem here for the QS parameter, since the learner will eventually run across the crucial evidence to change it. In this sense, the default setting is safe. But it is not safe for the other parameters. Recall that the parameter that assigns main stress looks for main stress in a foot-sized window at the edge of a word. Now, when it encounters a word like *qúmmin,* it applies the incorrect setting of the QS parameter to discover, incorrectly, that main stress is on the left (30a), not realizing that the whole word comprises only one foot (30b). Many variations on this theme can be produced, all of which show that incorrect default settings can be deadly to dependent parameters.[12]

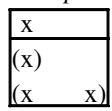
What recourse does an incremental learner have in such situations? To keep from making a false move in setting the main stress parameter, it has to be sure that its values for the QS

[12] For similar reasons, the Subset Principle (Berwick 1985) cannot ensure that a learner will not mistakenly become trapped in a superset value of a parameter for which a subset-superset relation exists.

(30) a. *Selkup: incorrect analysis*




b. *Selkup: correct analysis*




parameter are correct. We might propose that it should hold off setting any parameters that depend on a parameter—say, QS—until it is sure about the setting of the QS parameter. This is feasible for a given parameter if it is in its marked state, because a change to a marked state is only triggered by positive evidence and because, since parameters are binary, no further changes will occur. But what if the language being learned has the default value of a parameter (e.g., what if it is really QI)? In that case no positive evidence will ever come to confirm the setting of the independent parameter.

One proposal, then, is that we set some time limit *t* for setting each parameter: if after time *t* there has been no positive evidence to move to the marked setting, then we freeze the default setting. But now the incremental learner has in fact become a batch learner. A second possibility, the one adopted here, is the following: allow the incremental learner to set parameters as before, but impose the principle that when a parameter changes its value, all parameters that depend on it must revert to default. No more refined procedure is possible, on the assumption that the learner does not remember why it set some parameter to a particular value. Even if it could remember the crucial forms, it would not be easy to unravel the reasoning that led to every change. In the case at hand, then, the learner could set the Selkup main stress (and other) parameters to various incorrect marked settings while the QS parameter is in its default state, but it would have to wipe these out as soon as the QS parameter changes.

Although this modification preserves the incremental model, it brings it closer to the batch learner, because until a parameter is at its correct setting, nothing the learner does with regard to parameters that depend on it really matters. If this model is correct, it has an interesting consequence for the developmental problem of acquisition: it provides a mechanism for creating superset errors in the course of acquisition that do not require negative evidence to retreat from.

### *2.2 Grammar Acquisition and the Growth of Complexity*

The proposal that parameter setting follows an ordered path has a natural connection to the literature that understands the development of a child's phonological system in terms of growth of complexity along several dimensions (Macken 1978, Waterson 1978, 1987, Dresher, forthcoming). The development of prosodic structure is commonly viewed in terms of syllabic and metrical templates that expand from simple to more complex. Markedness can also be viewed in terms of complexity.[13] Jakobson (1941/1968) proposed that distinctive features develop by a series of

[13] For various elaborations of this idea, see Avery 1996, Dresher and van der Hulst 1995, 1997, van der Hulst 1994, Kaye, Lowenstamm, and Vergnaud 1985, and Rice 1992.

binary splits that take place in order of increasing markedness of the relevant features, an approach pursued by Jakobson and Halle (1956), Ingram (1989), and Dinnsen (1992), among others. Thus, the development of segmental inventories can be understood, like the development of prosodic structure, in terms of growth of complexity (Rice and Avery 1995). I assume that other areas of grammar can also be analyzed this way.

If this general conception is correct, it follows that, at least for some parameters, the choice between a default and marked setting is a choice between a less complex and more complex grammar. As we shall see, a number of other learning algorithms posit that the space of possible grammars is uniform, in the sense that the learner could just as easily happen upon one grammar as another. On the view taken here, the space of grammars is not uniform in this way, but consists of regions of differing degrees of complexity. It is expected, and the evidence suggests, that the series of grammars a learner traverses in the course of acquisition does not resemble a random succession of states, but represents a movement from lesser to greater complexity along a number of dimensions.

I now turn to consider some other learning algorithms that have recently been proposed. I think they all illuminate various aspects of the learning problem; but each makes a crucially wrong assumption about the nature of this problem.

## 3 The Triggering Learning Algorithm (Gibson and Wexler 1994)

Let us consider first the model sketched by Gibson and Wexler (1994).[14] Gibson and Wexler formulate a general scheme they call the Triggering Learning Algorithm (TLA), described in (31).

(31) *The Triggering Learning Algorithm (Gibson and Wexler 1994:409–410)*
Given an initial set of values for *n* binary-valued parameters, the learner attempts to syntactically analyze an incoming sentence *S*. If *S* can be successfully analyzed, then the learner's hypothesis regarding the target grammar is left unchanged. If, however, the learner cannot analyze *S*, then the learner uniformly selects a parameter *P* (with probability 1/*n* for each parameter), changes the value associated with *P*, and tries to reprocess *S* using the new parameter value. If analysis is now possible, then the parameter value change is adopted. Otherwise, the original parameter value is retained.

This algorithm incorporates two constraints, the Single Value Constraint (32) and the Greediness Constraint (33).[15]

[14] See also Berwick and Niyogi 1996, Fodor 1998, Frank and Kapur 1996, and Kapur 1994 for critical discussion, refinements, and further investigation of the properties of this model. Fodor (1998, in press a,b) has been developing a learning model that, in its emphasis on structural cues rather than surface sentences, has a certain affinity with the cue-based learner; however, it is not possible to enter into a discussion of this model here.

This section has benefited from insightful comments on a previous draft by Janet Fodor.

[15] These constraints are due to Robin Clark (1989, 1990; he does not, however, accept them as being valid). Frank and Kapur (1996:627) propose that these constraints should be more transparently called the Adjacency Constraint and the Constant Progress Constraint, respectively.

(32) *The Single Value Constraint*
Assume that the sequence $\{h_0, h_1, \ldots, h_n\}$ is the successive series of hypotheses proposed by the learner, where $h_0$ is the initial hypothesis and $h_n$ is the target grammar. Then $h_i$ differs from $h_{i-1}$ by the value of at most one parameter for $i > 0$.

(33) *The Greediness Constraint*
Upon encountering an input sentence that cannot be analyzed with the current parameter settings (i.e., is ungrammatical), the language learner will adopt a new set of parameter settings only if they allow the unanalyzable input to be syntactically analyzed.

The description of the TLA in (31) contains a crucial equivocation: when Gibson and Wexler state that ''*S* can be successfully analyzed'' by the learner, what they mean is that the learner can match the surface form of *S*, not that the learner can analyze *S* correctly (in other words, the TLA learner is concerned with weak, not strong, generative capacity). It is a guiding assumption of the TLA that the learner does not know if its grammar, or any part of its grammar, is correct or not.

The notion of trigger implicit in (31) is the following: a trigger is a sentence that the learner is not able to analyze in its current grammar, but is able to analyze (correctly or incorrectly) in an adjacent grammar. This operative notion of trigger is different from either a global or local trigger as defined by Gibson and Wexler.

(34) *Triggers (Gibson and Wexler 1994:409)*
a. A *global trigger* for value $v$ of parameter $P_i$, $P_i(v)$, is a sentence *S* from the target grammar *L* such that *S* is grammatical if and only if the value for $P_i$ is $v$, no matter what the values for parameters other than $P_i$ are.
b. Given values for all parameters but one, parameter $P_i$, a *local trigger* for value $v$ of parameter $P_i$, $P_i(v)$, is a sentence *S* from the target grammar *L* such that *S* is grammatical if and only if the value for $P_i$ is $v$.

Put informally, a global trigger is a sentence of the target language that requires the learner to set one parameter to its correct value, wherever in the parameter space the learner is; a local trigger is a sentence of the target language that requires the learner at a particular space to set one parameter to its correct value. Such triggers, which point the learner in the right direction and not a wrong one, would be very useful to a learner that could identify them; however, the TLA contains no mechanism by which a learner could distinguish such triggers from ''false'' triggers (sentences that cause the learner to make an incorrect change in its grammar), and hence they play no role in the operation of the TLA.

An example of how this learning algorithm is supposed to work is given in figure 1, where each square represents a setting of two syntactic parameters. The first parameter determines whether the head of [Spec, X′] is initial (value 1) or final (0). In this case the head is the verb (*V*) and its specifier is the subject (*S*). The second parameter encodes whether the head of a complement is initial or final, here exemplified by the relation between a verb and its object (*O*). These two parameters define a space with four states.

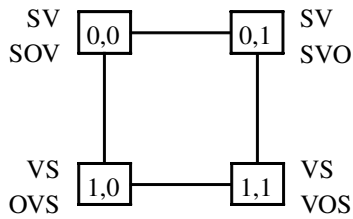



**Figure 1**
Parameter space: (specifier-head final/initial, complement-head final/initial): final = 0, initial = 1

Assume now that the target language is VOS (1,1) and that the learner's current hypothesis is SOV (0,0). Suppose the learner hears a sentence of the form *V O S*. This sentence is not parsable by the learner, who now determines that the current state is not correct. Even though there is only one setting of parameters that corresponds to *V O S,* the learner would have to change both parameters to reach it. This is not allowed by the Single Value Constraint, which makes available only the two neighboring spaces. Neither space yields the target *V O S*. Therefore, according to the Greediness Constraint, the learner cannot move. Thus, the sentence *V O S* is not a trigger to a learner at (0,0).

Fortunately, in this case the target includes another type of sentence that the learner will eventually hear, namely, *V S*. *V S* is a trigger to a learner at (0,0), since there is a neighboring space that parses it, namely, (1,0). So the learner moves to there. From there, a further presentation of *V O S,* which is a trigger to a learner at (1,0), will take the learner to the target.

Gibson and Wexler point out that the TLA will not be successful in the case of subset parameters, that is, parameters for which the sentences generated under one value are a proper subset of the sentences generated under the other value; in that case the learner who is mistakenly in the superset state will have no triggers, since all input sentences can be analyzed. Gibson and Wexler restrict their discussion to nonsubset parameters.

The main point of their article is that the TLA does not guarantee that a learner will converge on the target, because there are nonsubset parameter sets for which there are no triggers. The type of example they illustrate involves local maxima, which are triggerless islands in the parameter space.

Their illustration requires us to add one more parameter, the parameter that is responsible for verb-second (V2) effects (assuming this is one parameter). This parameter has the value 0 if the grammar is not V2, and 1 if it is. V2 has the potential to obscure the effects of the other parameters by requiring the verb to move into second position, and some other constituent into first position. This parameter space can be diagrammed as in figure 2.

Suppose the target is (0,1,0): SVO with no V2. Such a language generates surface strings like those represented schematically in (35a). Suppose also that the learner is currently at (1,1,1): VOS + V2, with forms as in (35b). There are some sentences that look the same in both, even

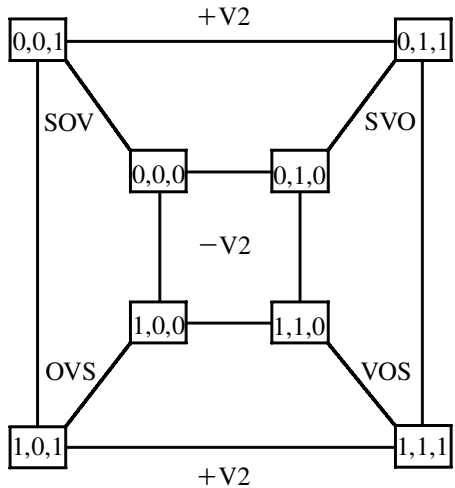



**Figure 2**
Parameter space adding V2: 0 = −V2, 1 = +V2

though their structures are different (e.g., *S V, S V O*). So a learner in (1,1,1) will not move when it encounters any of these. It turns out that all the potential triggers (35c) are not in spaces accessible to the learner. For example, the target string *Adv S V* is not parsable by the learner; but none of the three moves the learner can make results in a grammar that parses this string. It would have to change two parameters to see any improvement. Therefore, the learner is stuck at a local maximum.

(35) *Sample structures: target (0,1,0), source (1,1,1) is local maximum*
   a. *(0,1,0):* S V, S V O, S V O1 O2, S Aux V, S Aux V O, S Aux V O1 O2, Adv S V, Adv S V O, Adv S V O1 O2, Adv S Aux V, Adv S Aux V O, Adv S Aux V O1 O2
   b. *(1,1,1):* S V, S V O, O V S, S V O1 O2, O1 V O2 S, O2 V O1 S, S Aux V, S Aux V O, O Aux V S, S Aux V O1 O2, O1 Aux V O2 S, O2 Aux V O1 S, Adv V S, Adv V O S, Adv V O1 O2 S, Adv Aux V S, Adv Aux V O S, Adv Aux V O1 O2 S
   c. *Potential triggers:* Adv S V, Adv S V O, Adv S V O1 O2, Adv S Aux V, Adv S Aux V O, Adv S Aux V O1 O2

Gibson and Wexler consider several ways of overcoming the problem of local maxima. They observe that local maxima arise when the learner mistakenly gets into a +V2 state and that the problem would not arise if the learner could be prevented from trying +V2 until it has tried −V2 options. The solution they appear to favor is to adopt default states for parameters together with requiring that parameters be set in a partial order. Thus, their model comes closer to the

cue-based learner in these respects. However, problems remain in this regard. It is not enough for the learner to begin in a −V2 state as default; it must be prevented from entering a +V2 state until it has had a chance to consider all the potentially correct −V2 states. So we can suppose that +V2 states are simply not available to the learner for a time. But for how much time? No matter what time span is chosen, the TLA cannot guarantee that a learner will converge on the correct −V2 parameter setting in that time. A cue-based learner has a built-in solution to this problem, because unlike the TLA triggers, the cues available to a cue-based learner are unambiguous. The learner remains in the default state (−V2, under the same simplifying assumptions made by Gibson and Wexler) until it sees a positive cue to move to the marked state. Nothing more needs to be added.[16]

One might suppose that the difficulties faced by the TLA here are due to some special properties of the V2 parameter. For instance, +V2 languages have more types of sentences than their −V2 counterparts (e.g., where SVO −V2 has the single word order *S V O*, any of the +V2 languages has both *S V O* and *O V S*). Thus, although −V2 languages are not strictly subsets of +V2 languages, they are significantly smaller; and recalling the TLA's difficulties with subset languages, it is reasonable to attribute some of its difficulties here to the imbalance in the number of sentences in −V2 and +V2 languages. However, the problem of local maxima is not confined to such situations, but arises also when all the languages in the learning space have the same number of target forms.

To see this, let us consider how the TLA fares with metrical parameters. The parameter set used in the following experiment is a reduced version of the one discussed in section 1. The parameters discussed earlier enter into a number of dependencies, where not all combinations of parameter values are realized. To cite just two examples: if feet are QI, no difference results from setting the QS parameter to one or another value; and unbounded feet are assumed to always be QS, so the combination of quantity insensitivity and unbounded feet is not available. Such dependencies are no trouble to a cue-based learner, which instead profits from not having to consider certain paths. It is not obvious how to treat redundant or inactive states in a TLA model, however.[17] To keep such situations to a minimum, therefore, the parameters in (36) were selected.

(36) *Some metrical parameters*
1. *Main stress (MS):* Main stress is on the {left/right}.
2. *Extrametricality (EM):* The final syllable {is not/is} extrametrical.
3. *Directionality (Dir):* Feet are constructed from the {left/right}.
4. *Headedness (Hd):* Feet are headed on the {left/right}.
5. *Syllable quantity (QI/QS):* Feet are {QI/QS}.
6. *Secondary stress (SS):* Nonbranching (light) feet {are/are not} stressed.

[16] It remains, of course, to determine what the cue is (or the cues are). A learner must be able to distinguish between positions that are tied to particular arguments and those that can be occupied by any constituent (such as the preverbal position in V2 languages). Such a determination cannot be made by trying to match surface sentences; instead, it requires the learner to keep track of patterns. See Lightfoot 1991:chap. 3, forthcoming, for discussion and some proposals.

[17] See Frank and Kapur 1996:651–652 for some discussion of this sort of situation.

The changes from the parameters discussed earlier are as follows:

1. Main stress is as before.

2. Extrametricality is limited to the right side (cf. the NONFINALITY constraint, Prince and Smolensky 1993:52).

3, 4. Only binary feet are considered here.

5. Syllable quantity is limited to only two choices: QI or QS. Only one type of quantity sensitivity is assumed, whereby any branching rime or nucleus is considered heavy.

6. Destressing of secondary stress is limited to just the following choice: feet consisting of a single syllable in a QI language, or of a single light syllable in a QS language, {do/do not} receive secondary stress. For example, English selects SS-N; the alternative choice would cause words like *agenda* to have secondary stress on the initial light syllable.

The above parameter set generates $2^6$ (64) languages. Each language was assigned 4 two-syllable words, 8 three-syllable words, 16 four-syllable words, and 10 five-syllable words. Thus, each language has 38 words. Four pairs of languages are extensionally equivalent: their surface stress patterns are identical, though their grammars assign different structures. Since a learner would have no evidence to decide which grammar is correct, these languages are excluded as target grammars from the following discussion.

An analysis of how the TLA would apply to the remaining 56 languages yields the results in table 3. In this table *local maxima* are, as described earlier, states (excluding the target itself) from which the learner cannot exit, and *cul-de-sacs* are states that do not connect to the target, though exit is possible to one or more dead-end states. A learner that arrives at any of these states is guaranteed to fail to reach the target. A *dangerous* state is a state that connects to a local maximum or cul-de-sac, as well as to the target. Although a learner in a dangerous state has a

**Table 3**
64-state Triggering Learning Algorithm

| | | Problem states | | |
|---|---|---|---|---|
| Number of targets | Safe states | Local maxima | Cul-de-sacs | Dangerous states |
| 2 | 16 | 4 | 8 | 36 |
| 2 | 40 | 13 | 9 | 2 |
| 2 | 42 | 3 | 1 | 18 |
| 2 | 48 | 1 | 11 | 4 |
| 2 | 48 | 1 | 1 | 14 |
| 2 | 48 | 0 | 6 | 10 |
| 2 | 52 | 12 | 0 | 0 |
| 8 | 55 | 6 | 0 | 3 |
| 6 | 55 | 9 | 0 | 0 |
| 2 | 56 | 2 | 0 | 6 |
| 26 | 64 | 0 | 0 | 0 |

chance of reaching the target, success is not guaranteed. In terms of the goal of a learning theory for language, all of these states are *problem* states. *Safe* states are states that do not connect to any problem states; assuming that each triggered transition from a safe state has some probability greater than zero, arrival at the target is guaranteed in the limit. In the case at hand, 26 languages have no problem states, whereas 30 languages have between 8 and 48 problem states. In other words, even though there are no subset relations in the data set, and all languages have the same number of words, nearly one half of the languages cannot be guaranteed to be learnable by the TLA.

To take a concrete example, consider a language like Maranungku, discussed above. In terms of the current parameter set, it is characterized as in (37a); the (schematic) words used in this exercise are given in (37b).

(37) *Language 1 (Maranungku)*
   a. *Parameters:* MS-L, EM-N, Dir-L, Hd-L, QI, SS-Y
   b. *Words:* L2L0, H2L0, L2H0, H2H0, L2L0L1, H2L0L1, L2H0L1, H2H0L1, L2L0H1, H2L0H1, L2H0H1, H2H0H1, L2L0L1L0, H2L0L1L0, L2H0L1L0, H2H0L1L0, L2L0H1L0, H2L0H1L0, L2H0H1L0, H2H0H1L0, L2L0L1H0, H2L0L1H0, L2H0L1H0, H2H0L1H0, L2L0H1H0, H2L0H1H0, L2H0H1H0, H2H0H1H0, L2L0L1L0L1, H2L0H1L0L1, L2L0L1H0L1, L2H0H1L0L1, H2H0H1L0L1, H2H0L1L0L1, H2H0L1L0H1, L2L0H1H0H1, L2H0H1H0H1, H2L0H1H0H1

In these forms *L* indicates a (potentially) light syllable, *H* indicates a (potentially) heavy syllable,[18] and the number following a syllable indicates its level of stress: main stress (2), secondary stress (1), or no stress (0). Computing all the possible transitions using the TLA, we find that when language 1 is the target, there are six local maxima and three dangerous states, totaling nine problem states from which transition to the target is not assured. An example of a local maximum is language 8.

(38) *Language 8 (like English/Latin, but QI)*
   a. *Parameters:* MS-R, EM-R, Dir-R, Hd-L, QI, SS-Y
   b. *Words:* L2L0, H2L0, L2H0, H2H0, L2L0L0, H2L0L0, L2H0L0, H2H0L0, L2L0H0, H2L0H0, L2H0H0, H2H0H0, L1L2L0L0, H1L2L0L0, L1H2L0L0, H1H2L0L0, L1L2H0L0, H1L2H0L0, L1H2H0L0, H1H2H0L0, L1L2L0H0, H1L2L0H0, L1H2L0H0, H1H2L0H0, L1L2H0H0, H1L2H0H0, L1H2H0H0, H1H2H0H0, L1L0L2L0L0, H1L0H2L0L0, L1L0L2H0L0, L1H0H2L0L0, H1H0H2L0L0, H1H0L2L0L0, H1H0L2L0H0, L1L0H2H0H0, L1H0H2H0H0, H1L0H2H0H0

A TLA learner trying to learn language 1, whose current grammar is that of language 8, would have no shortage of evidence that its grammar is incorrect. Except for the four bisyllabic words,

[18] Syllables are only potentially light or heavy, because the distinction is relevant only in QS languages. In QI languages the designations *L* and *H* do not reflect the analysis assigned by the grammar (for which all syllable types are equal). They are retained here to facilitate crosslinguistic comparisons.

**Table 4**
State 8 is a local maximum for state 1

| | Analysis | Surface |
|---|---|---|
| Word to match from language 1 | (tá ta)(tà) | tátatà |
| Word derived by current grammar of language 8 | (tá ta)⟨ta⟩ | *tátata |
| a. Change main stress to left | (tá ta)⟨ta⟩ | *tátata |
| b. Change extrametricality to no | (tà)(tá ta) | *tàtáta |
| c. Change direction to leftward | (tá ta)⟨ta⟩ | *tátata |
| d. Change foot head to right | (ta tá)⟨ta⟩ | *tatáta |
| e. Change quantity to QS | (tá ta)⟨ta⟩ | *tátata |
| f. Change secondary stress to no | (tá ta)⟨ta⟩ | *tátata |

which happen to coincide in the two languages, there is no word in the target that can be assigned an analysis in the current grammar; that is, 34 of 38 words can potentially inform the learner in language 8 that its grammar is wrong. However, it is unable to profit from this information because no word can be remedied by making just one change in the grammar.

For example, the target word L2L0L1 (say, a word like *tátatà*) is not generable by the grammar of language 8, which instead generates L2L0L0 (*tátata*). These words look fairly close, but no single change will enable the grammar of language 8 to produce the required pattern, as shown in table 4. This example highlights the extent to which a surface form is a composite, made up of diverse interacting factors. There is no reason to expect that there will always be an adjacent grammar, one parameter change away, that will have the required forms.

The existence of such ''stepping stones'' is due to no principled considerations. For example, language 8 is one of six local maxima for language 1. When language 8 is itself the target, it has no local maxima, but only because of a lucky accident. It can be shown that there are eight states from which the learner cannot exit by making a correct change in its parameter set. In those cases, however, there is an *incorrect* change the learner can make that moves it to an adjacent state further from the target; from that further state, routes to the target can be found.

The principle of strategic retreat has indeed been observed to play a role in language learning, but not in this way. For example, it is well known that when children first acquire a rule, they sometimes overgeneralize it for a time, even to the extent of replacing correct forms with incorrect ones: hence the familiar sequence *went* > *goed* > *went*. In these cases, though, the surface forms become worse as a consequence of the grammar's becoming better. The cue-based learner exhibits similar sequences: because it depends on its cues and is not trying to directly match surface forms, it will make a change in its grammar even if the immediate result is that some output forms move further from the target. The TLA learner operates the other way around: to match a surface form, the TLA learner is as ready to retreat from the target grammar as it is to progress. Moreover, the matching of surface forms is entirely superficial and may conceal drastic differences in analysis.

Because the TLA learner will move to any adjacent state that appears to match an input form, we might expect it to do a good deal of ''thrashing'': moving back and forth between

states, perhaps indefinitely. Thrashing is mentioned as a theoretical possibility by Frank and Kapur (1996), but it is a very real prospect in the parameter set we have been considering. For example, a learner trying to acquire the English/Latin pattern discussed earlier might have every parameter correct except quantity sensitivity. Therefore, it would correctly parse a word like *Cánada* as (*Cá na*) ⟨*da*⟩, but it would incorrectly parse *agénda* as (*á gen*)⟨*da*⟩. Instead of changing the syllable quantity parameter, the learner might instead try to revoke extrametricality, producing the correct-sounding (*a*) (*gén da*). Now the previously correct *Cánada* has become unparsable—it would now be analyzed as (*Ca*)(*ná da*)—a fault that could be corrected by restoring extrametricality, returning to the previous state.

Thrashing can involve a large number of states, again depending on accidental properties of surface forms. To put some numbers to this: in the space defined by the six metrical parameters in (36), for target languages that have no problem states, the number of retrograde transitions (away from the target) ranges from 19.6% to 37.2% of all transitions. When the target is language 8 (38), for example, 22.6% of all possible transitions are away from the target.[19] As long as there is no bias to move in certain directions rather than others, convergence to the target in the limit is still guaranteed (the target being the only local maximum).[20] However, learnability in the limit is a very weak criterion of adequacy, and the routes taken by a TLA learner on the way to the target resemble no known observed acquisition sequence.

Over and above the technical difficulties, the TLA runs into some serious conceptual problems. The essential difference between the TLA and the cue-based learner has to do with the conception of what the learner is trying to do, and what constitutes a trigger, or cue. Under the TLA, the learner is trying to match the target input forms; hence, a trigger is an actual input form. A cue-based parameter learner, by contrast, is not trying to match the target forms; instead, it uses them as sources of cues. Thus, whereas triggers in the TLA are extensional entities (actual words or sentences that are part of E-language), cues are intensional entities. Similarly, the two learning models treat the notion of a learning path in different ways. In the cue-based model, the path is fixed by UG; in the TLA model, learning paths arise purely as a result of accidental features of the input.

Further, Gibson and Wexler's account is predicated on the assumption that the target sentences come in the form of strings like those in (35), which have the form *S O V, Adv Aux S O V,* and so on. Of course, the real target sentences that the learner sees are not in that form, but

[19] This figure does not necessarily equal 22.6% of all possible moves, since this calculation does not count how many input strings trigger each move. Suppose, for example, that there are two transitions out of a given state: a move toward the target triggered by six possible input forms, and a move away from the target triggered by four forms. The number of possible transitions away from the target is 50% of all the transitions from that state, but the probability of making a retrograde transition from there is 40%, assuming that all forms are equally probable and that there are no other biases in the system. See Frank and Kapur 1996 and Berwick and Niyogi 1996 for further discussion of stochastic and nonstochastic versions of the TLA.

[20] Bias in choosing parameters to change could help the learner in some cases and be deadly to it in others. For instance, a bias to always try to change extrametricality before tampering with syllable quantity would convert the example discussed above from a case of thrashing, where the learner will eventually try to change the syllable quantity parameter, to a potential cul-de-sac with no route to the target. In general, I do not see any way to build biases into a TLA that will be guaranteed to not be harmful, unless the model becomes considerably more like the cue-based learner.

**Table 5**
Representations of *Je le vois*

| Subject | Object | Analysis |
|---|---|---|
| a. nonclitic | nonclitic | *S O V* |
| b. nonclitic | clitic | *S V* |
| c. clitic | clitic | *V* |

are actual utterances: for example, *John kicked the ball, Je le vois* '(lit.) I it see (I see it)'. A successful analysis of a complete sentence involves not just its syntactic word order, but everything else as well: phonology, morphology, and so on. So the parameters in play are not just those affecting word order, but all of them. Now, chances are that a learner, especially at an early stage, is unable to match even simple sentences with respect to every component of the grammar: not just word order may be off, but also morphology, inflection, segmental phonology, metrical and prosodic properties, and so on. So if a learner hears a sentence of the form *S V O* and is currently at SOV, a change to SVO will still not result in a complete match for the whole sentence. Similarly, any change in another type of parameter—say, a morphological one—might result in a successful match there; but the learner will not consider it a success, because the word order is still not right. Recall that a learner does not know what effect any given parameter has, and is not satisfied with improvements that fall short of success. Taken literally, then, the TLA would not let a learner get off the ground. This is because it requires a chain of complete successes. Even in a single domain, there is no guarantee that such a chain could be compiled; over the grammar as a whole, this goal appears to be out of reach, for no target is small enough to be perfectly matched, especially at early stages.

Let us suppose, then, that Gibson and Wexler intend that the learner can separate the word order properties of a sentence from its other properties. Let's say that success must be total only within this domain. The problem with this is that the domain of facts influencing the setting of word order parameters is not limited to word order. Suppose that pronouns can be clitics, or not. *Je le vois* could then be an example of *S O V* (if the subject and object are not clitics), or *S V* (if the object is a clitic, so that there is no lexical material in the actual object position), or just *V* (if both subject and object are clitics), as illustrated in table 5.

The learner's analysis depends on the current state of its grammar.[21] The terms *S, V, O* are not primitives coming from the target, but are assigned by the learner, based on knowledge of the grammar. Therefore, we cannot limit the parameter space relevant to word order only to word order parameters. For example, if the learner is currently assuming SVO plus (a) of table 5 and

[21] An actual example illustrating this point is provided by Lightfoot (1997, forthcoming). It has been argued that, at a certain stage in the history of Middle English, southern dialects treat subject pronouns as clitics (van Kemenade 1987), but northern dialects do not (Kroch and Taylor 1997). Thus, a sequence *subject pronoun – XP – finite verb* is consistent with a V2 analysis in the south, but not in the north. See also Clark and Roberts 1993:338 on some possible word-order confusions created by the uncertain status of preverbal subject pronouns in the history of French.

hears the sentence *Je le vois,* it perceives the sentence as *S O V*. Now the learner can change word order and move to SOV plus (a) of table 5; or, without changing word order, it can move to SVO plus (b) of table 5. Clearly, word order parameters cannot be correctly set without taking into account clitic status and other such matters. But how does the learner know which group of parameters forms a subspace within which matching must be perfect? It appears that, even on Gibson and Wexler's own account, the learner must have some idea about what sort of thing a parameter does.

## 4 A Genetic Algorithm (Clark 1990, 1992, Clark and Roberts 1993)

I would now like to look briefly at another approach to parameter setting developed by Clark (1990, 1992) and applied to V2 changes in the history of French by Clark and Roberts (1993).

Contrary to the approach taken here, Clark does not believe it is possible to associate reliable cues to parameters. Rather, he proposes to assign a fitness metric that gives the relative fitness of a grammar compared with others. His idea is that parameter setting proceeds by way of a genetic algorithm that enacts a Darwinian competition of survival of the fittest. He proposes that a learner simultaneously considers a number of competing hypotheses. Initially, these hypotheses may be selected randomly. Each candidate is exposed to input that it attempts to parse. At the end of a round of parsing, the learner assesses how well each candidate did. The candidates are ranked according to their relative fitness. The fittest go on to reproduce candidates in the next generation; the least fit die out. Through successive iterations of this procedure, the candidate set presumably becomes increasingly fit and converges toward the correct grammar.

This approach is at the opposite pole from the cue-based model. The cue-based learner knows why it sets a particular parameter to a particular value—because it sees or fails to see a cue—but it has no way to evaluate the overall success of its grammar. The learner following the genetic algorithm has no idea what contribution any particular parameter makes, but it has an exquisite sense of the overall relative success of the grammar.

The proposed fitness metric is given in (39).

(39) *Fitness metric (Clark 1992, Clark and Roberts 1993)*

$$\frac{\left(\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j + c\sum_{j=1}^{n} e_j\right) - (v_i + bs_i + ce_i)}{(n-1)\left(\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j + c\sum_{j=1}^{n} e_j\right)}$$

where

$v_i$ = the number of violations signaled by the parser associated with a given parameter setting;

$s_i$ = the number of superset settings in the counter; $b$ is a constant superset penalty $< 1$;

$e_i$ = the measure of elegance (= number of nodes) of counter $i$; $c < 1$ is a scaling factor.

There are three main terms in the metric. The first term, $v$, refers to the number of violations in parsing an input sentence signaled by the parser associated with a given parameter setting. Whereas in the TLA the learner is told only if a hypothesis succeeds or fails, Clark proposes to quantify the failure in terms of the number of violations incurred. The sum term totals all the violations created by all the candidates. Suppose there are five candidates that together total 50 violations. We then subtract from the total the number of violations incurred by any candidate $i$ and divide by the total (multiplied by $n-1$); this gives us a measure of how well candidate $i$ is doing compared with the rest. For example, if candidate 1 creates 10 violations, its score is $50 - 10 = 40$ divided by some number; if candidate 2 creates 30 violations, its score is $50 - 30 = 20$ divided by that number, a lower score.

This term is the main component of the fitness metric. Clark builds in two other terms, scaled down by constant factors to make sure they are small relative to the $v$ term. The second term is a superset penalty, designed to have the effect of the Subset Condition. If two candidates differ only in one subset parameter, and the target language is the subset language, they ought to score identically with respect to violations, since anything that the subset parameter value can parse, the superset value can parse, too. To keep the learner out of the superset, Clark builds in a penalty, the term $s$. So if two candidates both have 10 violations, they will have equal scores of 10 (roughly, forgetting about the subtraction and division). If candidate 1 has one superset parameter value, its score will be lowered by the constant term $b$. Candidate 2, let's say with two supersets, is penalized by $2b$. Clark (1990) suggests that $b$ is very small, around 0.00002: it has to be much smaller than 1, since it should not count nearly as much as a violation. Whatever the number, it is enough to put candidate 1 ahead of its superset competitor. The third term, $e$, is another refinement, a measure of elegance, which Clark roughly equates with the number of nodes that a candidate hypothesis needs to parse the target sentences. This is to give the effect of economy, preferring simple grammars to more complex ones. Clark and Roberts argue (1993:342) that empirical facts from the history of French show that the constant $c$ is greater than $b$: elegance counts more than subsetness.[22]

Although it may be instructive to experiment with algorithms of this type, there are several grounds for questioning the feasibility and plausibility of the proposed fitness metric that is the heart of the model. Consider first the subset penalty. This penalty refers to E-language (extensional) subsets, actual subsets calculated over sentences. Clark suggests that superset parameters are listed in a table; that is, they are supplied to the learner by UG. The learner following the genetic algorithm has no idea about what any individual parameter does; yet it does know which parameters create extensional supersets, an apparently paradoxical conclusion that, moreover,

[22] However, Clark and Roberts's historical account does not support their learning algorithm. In essence, their proposal is that V2 order in French was lost after the introduction of sentences of the form *XP Subject V,* which are incompatible with a V2 analysis. What needs to be explained is how such forms could be introduced into a V2 grammar in the first place. It appears that either the grammar had already lost V2, at least for some significant portion of speakers, or the incompatible forms were originally consistent with a V2 grammar (perhaps as in Middle English, in the case of subject pronouns that may have originally been clitics—see footnote 21), but then underwent a reanalysis. Though Clark and Roberts do not take a clear position on what caused the change, the possible causes they consider are all entirely compatible with a cue-based account.

**Table 6**
Effects of parameter settings: Selkup

| Parameters correct | | Words correct | | Syllables correct | | Main stress correct | |
|---|---|---|---|---|---|---|---|
| a. 4/10 | 40% | 2/8 | 25% | 7/20 | 35% | 3/8 | 37.5% |
| b. 6/10 | 60% | 1/8 | 12.5% | 7/20 | 35% | 5/8 | 62.5% |
| c. 7/10 | 70% | 4/8 | 50% | 12/20 | 60% | 4/8 | 50% |
| d. 8/10 | 80% | 5/8 | 62.5% | 14/20 | 70% | 5/8 | 62.5% |
| e. 9/10 | 90% | 5/8 | 62.5% | 14/20 | 70% | 5/8 | 62.5% |
| f. 9/10 | 90% | 3/8 | 37.5% | 10/20 | 50% | 3/8 | 37.5% |

encodes an extensional relation into UG. In the cue-based model, by contrast, subset languages are learnable to the extent that the learner has appropriate built-in (I-language) defaults. The learner need not—and cannot—know about extensional subset relations.

Second, like the TLA, the proposed genetic algorithm allows for no orderly progression in the learner's developing grammar. Candidate parameter settings are let loose over the entire parameter space, making it just as likely that a given learner will entertain a very complex grammar at any point in development as a simple one, entirely independently of the input data.[23] However, progressions from complex to simple grammars are not observed in actual acquisition.

Third, it is not clear whether a useful fitness metric can be devised for every aspect of the grammar. With respect to metrical parameters, for example, there is no clear correlation between the number of words correct and the distance from the target. Imagine a language with simple alternating stress. If we change the foot parameter from trochee (left-headed) to iamb (right-headed), every syllable will receive the wrong stress. If we then move further from the target by changing other parameter values in the wrong direction, performance—in terms of syllables or words correct—will appear to improve. In general, depending on the situation, small changes can have big effects and big changes can have small effects. In other words, there does not appear to be a ''smooth'' relation between distance in parameter space (the intensional grammar) and distance in terms of the number of (extensional) forms correct.[24] Therefore, any fitness measure based on purely extensional properties is liable to be an unreliable guide to the target grammar.

A concrete example will serve to illustrate what the learner is up against. Imagine a learner attempting to acquire the stress system of Selkup (14) using a genetic algorithm. Using the Dresher and Kaye 1990 system of parameters, I generated some random parameter settings and investigated what relative score a fitness metric might give them when applied to eight representative words. Since it is not obvious what criterion the fitness metric should use, I tried three different criteria: words correct, syllables correct, and main stress correct. Some results are given in table 6.

[23] I leave aside the question of what grammar a learner can be said to be entertaining in the presence of any number of competing candidates.

[24] For more on smoothness, see Berwick and Niyogi 1996:612, 614 and Frank and Kapur 1996:644, as well as footnote 26.

**Table 7**
Effects of parameter settings: target language 1

| Parameters correct | Words correct in each state | Average correct | Median state | Most states |
|---|---|---|---|---|
| 0 | 4 | 4 | 4 | 4 |
| 1 | 1, 4, 4, 4, 4, 8 | 4.2 | 4 | 4 |
| 2 | 0, 0, 1, 2, 4, 4, 4, 4, 4, 4, 4, 5, 8, 8, 12 | 4.3 | 4 | 4 |
| 3 | 0, 0, 0, 0, 2, 2, 4, 4, 4, 4, 4, 4, 4, 6, 6, 8, 8, 8, 8, 20 | 4.8 | 4 | 4 |
| 4 | 0, 0, 2, 3, 4, 4, 4, 4, 4, 4, 8, 8, 12, 18, 20 | 6.3 | 4 | 4 |
| 5 | 0, 4, 12, 20, 20, 20 | 12.7 | 16 | 20 |

As can be seen, none of the plausible potential criteria for a fitness metric show a reliable correlation with number of parameters correct. Although candidates (e) and (f) are each correct in all but one parameter (type of quantity sensitivity for (e), and main stress for (f)), they differ greatly in their apparent fitness, as reflected in their scores. Candidate (e) scores relatively high, as we might expect, though it has the same scores as (d), which is actually further from the target. More dramatically, candidate (f), with only one parameter wrong, scores worse in every category than (c), which has three parameters wrong. Similarly, (a), with only four correct parameters, has more words correct than (b), which has six. Moreover, these results can be influenced in unpredictable ways by the chance occurrence of various types of words, as well as by the nature of the target parameter set.

These conclusions are supported by an investigation of correlations between parameters correct and words correct on a larger scale, looking at the 64-state parameter space generated by the six metrical parameters discussed above with respect to the TLA. A typical result is that obtained when language 1 (Maranungku) is the target; see table 7. When we look only at the number of words correct in each state that has a given number of parameters correct, table 7 gives three different measures that may be relevant to a genetic algorithm. For example, there are twenty states that have three parameters correct. They range in numbers of words correct from 0 to 20. The average number of words correct is 4.8; the median state has 4 words correct; and the most frequently found number of words correct is also 4: seven states have 4 words correct, four states have 0, four have 8, two have 2 and 6, and one has 20.

In this case, at least the average number of words correct rises monotonically as the grammar comes closer to the target, but even this result cannot be guaranteed to hold for all targets in all parameter spaces. Of the 64 possible targets in this experiment, 16 have average numbers of words correct that are strongly nonmonotonic (i.e., the average number of correct words falls as the number of parameters correct rises, in some part of the space), and another 8 have weakly monotonic averages (i.e., the average number correct remains the same as the number of parameters correct rises, for at least one transition). An example of the former is target language 27 (parameters: MS-L, EM-R, Dir-L, Hd-R, QS, SS-Y); see table 8.

**Table 8**
Effects of parameter settings: target language 27

| Parameters correct | Words correct in each state | Average correct | Median state | Most states |
|---|---|---|---|---|
| 0 | 8 | 8 | 8 | 8 |
| 1 | 0, 4, 4, 4, 4, 14 | 5 | 4 | 4 |
| 2 | 0, 0, 2, 2, 3, 4, 4, 4, 4, 4, 6, 8, 8, 14, 18 | 5.4 | 4 | 4 |
| 3 | 0, 0, 0, 2, 3, 4, 4, 4, 4, 6, 6, 7, 8, 8, 8, 8, 9, 10, 11, 17 | 6 | 6 | 4/8 |
| 4 | 0, 0, 0, 4, 7, 8, 8, 8, 8, 10, 11, 12, 14, 16, 24 | 8.7 | 8 | 8 |
| 5 | 0, 8, 15, 18, 27, 33 | 16.8 | 16.5 | — |

Excluding the target itself, there are 17 states that look better than the state that has no parameters set correctly, and 34 states that look worse.

Thus, there is at best a very weak correlation between closeness to the target and outward success in terms that can be measured by a fitness metric.[25] This disparity between intension (the parameter set) and extension (the set of forms generated by a parameter set) is what gives rise to the Credit Problem in the first place. If there were a smooth curve connecting the two, we could use it to finesse the Credit Problem: though a learner would not necessarily know at any given point which parameters were wrong, it would have a good idea of how many were wrong, in which knowledge lie the seeds of a successful learning strategy. As things stand, however, a learner navigating by an extensional fitness metric is using a faulty compass. Neither Clark (1992), nor Clark and Roberts (1993), nor anyone else, to my knowledge, has shown that convergence to the target is guaranteed in such conditions.[26]

[25] Of course, it would be quite surprising if there were no connection at all between how close the grammar is to the target and how good the outward performance of the learner appears to be. The issue is whether the correlation is such as to make a metric of goodness-of-fit a reliable guide to a learner. For example, the correlations between number of parameters correct and number of words correct in languages 1 and 27 are statistically significant: it can be calculated that 12% of the variance in the data in language 1, and 15% in language 27, is accounted for by the relation of these two factors. This statistical significance, however, is of dubious value to the learner, since by the same token 88% and 85% of the variance, respectively, remains unaccounted-for ''noise.'' Thanks to Ron Smyth for help with these calculations and discussion of their significance.

[26] Turkel (1996) investigates the smoothness of a parameter space consisting of six metrical parameters that differ in some respects from the ones investigated here. He defines the region of parameter space around a target language, *I*, to be smooth iff the average score attained by the immediate neighbors of *I* with respect to some measure is higher than the average score attained by the other states. Using two different measures, he finds that the average score of the immediate neighbors of the target language (languages with five parameters correct) is indeed higher than the average score of the other states, and so the space around the target language is smooth by his definition. The same definition would find the parameter spaces we have investigated here to be smooth, also. However, it needs to be shown that this particular definition of smoothness, which computes an average score in a particular region of the parameter space, is relevant to the operation of the learning algorithm.

## 5 The Robust Interpretive Parsing/Constraint Demotion Algorithm (Tesar and Smolensky 1996, 1998)

Finally, I would like to consider the Robust Interpretive Parsing/Constraint Demotion (RIP/CD) learning algorithm proposed by Tesar and Smolensky (1996, 1998) for learning how to rank constraints in Optimality Theory. As opposed to a principles-and-parameters framework, Optimality Theory (OT; Prince and Smolensky 1993) posits that grammars consist of a common set of violable constraints that have language-particular rankings; lower-ranking constraints may be violated to preserve higher-ranking ones.

Tesar and Smolensky propose that the learning problem can be decomposed into two separate subproblems. The first is the problem, addressed by the Constraint Demotion algorithm, of ranking constraints given the optimal surface parse; the second problem, which remains to be solved, is that of arriving at the optimal surface parse. Tesar and Smolensky propose that the learner's developing grammar itself can be used, in tandem with constraint demotion, to gradually refine the learner's parses of the target forms, through a process they call Robust Interpretive Parsing (RIP).[27]

Tesar and Smolensky (1996) illustrate the procedure using the same English/Latin example we have been considering throughout. They assume that the learner must rank the following constraints that govern quantity sensitivity and extrametricality:[28]

(40) *Metrical constraints to be ranked (Tesar and Smolensky 1996)*
  a. *BISYLL:* A foot is bisyllabic.
  b. *WSP:* A heavy syllable is a head of a foot (Weight-to-Stress Principle).
  c. *PARSE-σ:* A syllable is parsed into a foot.
  d. *NONFIN:* A foot is not final in the prosodic word.

In this problem, the relative ranking of the first two constraints determines if stress is QI or QS: if BISYLL dominates WSP, a sequence *H H* or *L H* must be parsed as a single trochaic foot, violating WSP and yielding a QI stress system; if WSP dominates BISYLL, then a heavy syllable may not be the unstressed member of a foot, resulting in a QS system. Similarly, constraints (40c) and (40d) form an antagonistic pair whose ranking determines whether extrametricality holds: if PARSE-σ dominates NONFIN, final syllables will be parsed into a foot, if possible (no extrametricality); if NONFIN dominates PARSE-σ, final syllables will be extrametrical.

Tesar and Smolensky suppose that the learner has already correctly ranked the constraints governing foot form and directionality, as well as the position of main stress.[29] In this example,

[27] A different formulation of the learning problem is presented in Tesar and Smolensky 1993; see Dresher 1996 for discussion.

[28] The constraints in (40) are based on Prince and Smolensky 1993. WSP is what I have been calling quantity sensitivity, and NONFIN does the work of extrametricality on the right side. The various rankings of these constraints give the results indicated in conjunction with other constraints not mentioned here.

[29] If the learning path discussed in section 1 is correct, it is unrealistic to assume that a learner could in fact correctly determine foot form and directionality before determining whether the language is QS or not. This consideration is not relevant to the current case in its role as an example of how RIP works, but it will be quite relevant to the ultimate workability of this algorithm, as we shall see.

**Tableau 1**
Learner's grammar picks the wrong candidate

| Input: *América* | PARSE-σ | NONFIN | BISYLL | WSP |
|---|---|---|---|---|
| a. ☞ (A me)(rí ca) | | * | | |
| b. (A)(mé ri)ca | *! | | * | |

recall that feet are trochaic, constructed from the right, and that main stress is on the right. Tesar and Smolensky assume further that the learner has incorrectly ranked the four constraints in (40) as in tableau 1.[30] This ranking gives, in our earlier parametric terms, quantity insensitivity and no extrametricality, which are both incorrect. When presented with a word like *América,* with four light syllables, the grammar in tableau 1 parses it as (*A me*)(*rí ca*). The learner knows this is wrong, because such a form would have the overt stress pattern **Amerí́ca,* which is plainly incorrect, no matter what its foot structure.

So far the learner is in the same position as a TLA learner (modulo the switch from parameters to ranked constraints) that has arrived at an incorrect grammar and receives an input form that is not correct in terms of the current grammar. This datum informs the learner that its grammar is not correct, but, again, as in the TLA model, it gives the learner no information about how the grammar is incorrect. The TLA learner, as we have seen, makes a wild guess about what the incorrect parameter might be, and sees if changing it will result in the correct form. The RIP learner is quite different in this respect. It does not make a guess about which way to go; rather, its current grammar gives it a direction, as follows.

The learner knows that the correct analysis of the current form (*América*) must be such as to assign main stress on the antepenultimate syllable. Therefore, it uses its current grammar to find the highest-ranked candidate that matches this overt stress pattern. This candidate happens to be (b) in tableau 1, (*A*)(*mé ri*)*ca.* In the current grammar this candidate is not optimal, because it violates PARSE-σ, and hence loses to (a), which does not. The learner then changes its grammar so as to make the former parse the optimal one. Tesar and Smolensky propose that this change be minimal and that it be achieved by Constraint Demotion. In this case PARSE-σ is demoted to the stratum just below NONFIN, yielding the grammar in tableau 2.

The new grammar now requires extrametricality, but stress is still QI. Tesar and Smolensky observe that the subsequent presentation of a word like *agénda* will trigger a further change to quantity sensitivity, by a similar process. The optimal form in the new grammar is (a), which yields the incorrect overt stress pattern, **ágenda.* Again, the learner locates the optimal candidate that respects the observed stress pattern; here, there are two that are tied, (b) and (c). The learner picks one of these (presumably at random) and changes the grammar so as to make this form

[30] Tesar and Smolensky do not indicate secondary stresses in their examples, but presumably every foot has at least a secondary stress; thus, the first syllable in candidate (b) in tableau 2 is stressless, and the first syllable of candidate (c) is stressed.

**Tableau 2**
Learner's revised grammar and second target word

| Input: *agénda* | NONFIN | PARSE-σ | BISYLL | WSP |
|---|---|---|---|---|
| a. ☞ ( gen)da | | * | | * |
| b. a(gén)da | | **! | * | |
| c. (a)(gén)da | | * | *!* | |

optimal. Whether (b) or (c) is picked, the result will be at least the demotion of BISYLL below WSP, effecting a change from QI stress to QS. The algorithm continues in this fashion until it reaches the correct grammar and no further changes are triggered.

Let us turn now to an analysis of the RIP/CD algorithm. Does the procedure of altering the grammar so as to select the optimal form that accords with the overt stress pattern guarantee that the learner will converge on the correct grammar? The answer is that it does not. Since the algorithm selects the optimal candidate that is consistent with the overt stress pattern, it will always demote the lowest-ranked constraint(s) that it can. Put informally, it is relatively easy to demote constraints that are already ranked low, and relatively difficult to demote constraints that are ranked high.[31] This attribute is helpful to the extent that high-ranked constraints merit their position. However, in the algorithm outlined by Tesar and Smolensky, constraints may be highly ranked for the wrong reasons as much as for the right reasons. A constraint that is incorrectly highly ranked may, under a variety of circumstances, become undemotable, with fatal consequences for the learner.

Consider a simple example. Suppose the target language is Maranungku, discussed earlier. Recall that Maranungku has a simple alternating stress pattern consisting of trochees starting from the left. Suppose that a RIP/CD learner has arrived at a grammar that is mostly correct—binary feet, all syllables parsed into feet, no extrametricality—except that it thinks that feet are iambs constructed from the right.[32] An odd-parity word like *lángkaràteti* is consistent with this analysis, so the learner will make no change when presented with such a word. However, words with even parity, like *wélepènemànta,* are not consistent with the current grammar. What the RIP/CD learner does next is determined by its current constraint ranking. If the constraints governing directionality and foot form are ranked very low, the learner will head in the right direction. But suppose the current grammar is as in tableau 3, where directionality and foot form are ranked more highly than PARSE-σ and NONFIN. The optimal form that matches the overt stress pattern is not (b), but

[31] In this connection it can be shown that, given the assumptions made about the grammar at the stage shown in tableau 1—binary left-headed feet constructed from the right, main stress right—a word like *América* can be parsed *only* if one assumes that the final syllable is extrametrical. In the terminology of Gibson and Wexler (1994), *América* is a local trigger, forcing the demotion of PARSE-σ. If there were any other way to obtain main stress on the antepenult in this word, PARSE-σ would not be demoted.

[32] In OT, directionality is governed by a constraint that requires all feet to be aligned as closely as possible to the left or right edge; see McCarthy and Prince 1993.

**Tableau 3**
Learning Maranungku, 1
a. Grammar before exposure to even-parity word

| Input: *wélepènemànta* | DIR-R | HEAD-R | PARSE-σ | NONFIN |
|---|---|---|---|---|
| a. ☞ (welé)(penè)(mantà) | | | | * |
| b. (wéle)(pène)(mànta) | | *! | | * |
| c. (wé)(lepè)(nemàn)ta | | | *! | |

b. Grammar after exposure to even-parity word

| Input: *wélepènemànta* | DIR-R | HEAD-R | NONFIN | PARSE-σ |
|---|---|---|---|---|
| a. (welé)(penè)(mantà) | | | *! | |
| b. (wéle)(pène)(mànta) | | *! | * | |
| c. ☞ (wé)(lepè)(nemàn)ta | | | | * |

(c), which obeys the higher-ranked constraints requiring iambs from the right, at the cost of violating PARSE-σ. The RIP/CD algorithm will now demote PARSE-σ below NONFIN, keeping iambs from the right and adding final extrametricality.

The new grammar (tableau 3b), however, can no longer generate the stress patterns of odd-parity words (tableau 4a); the best candidate that matches the overt stress pattern is candidate (c) in tableau 4, which leads to the demotion of NONFIN below PARSE-σ. The result (tableau 4b) is that the learner is back to the grammar it began with. It is caught in an infinite loop, for it will always prefer to demote one of PARSE-σ or NONFIN, thus turning extrametricality on and off repeatedly, rather than tamper with foot directionality or headedness.

This is the RIP/CD counterpart of thrashing, discussed above in connection with the TLA. Recall that thrashing, though undesirable on various grounds, is less of a threat to a stochastic TLA (which will, eventually, choose a route that will enable it to exit from the loop) than it is to a deterministic or biased TLA. The RIP/CD algorithm is comparable to a deterministic TLA in this respect: it cannot escape from vicious loops of the sort described.[33]

The Achilles' heel of this algorithm is that the learner uses its current grammar to favor

[33] For further discussion of nonconvergence in the RIP/CD algorithm, see Tesar 1997, 1998. Tesar (1997) observes that the algorithm performs best when it is given a starting hierarchy WSP ≫ PARSE ≫ HEAD-R ≫ HEAD-L ≫ all other constraints. Note that this hierarchy resembles the learning path in (28), where syllable quantity (= WSP) precedes extrametricality (similar to PARSE) and where the parameters for headedness and directionality are set further down the learning path. Thus, it appears that this particular hierarchy works best because it reflects dependencies that are encoded in the learning path. Adopting a starting hierarchy of this kind brings the RIP/CD algorithm closer to the cue-based learner in this respect. It has not been shown, however, that any starting hierarchy can guarantee convergence of the RIP/CD algorithm in all cases.

**Tableau 4**
Learning Maranungku, 2
a. Grammar before exposure to odd-parity word

| Input: *lágkaràtetì* | DIR-R | HEAD-R | NONFIN | PARSE-σ |
|---|---|---|---|---|
| a. ☞ (langk)(ratè)ti | | | | * |
| b. (lngka)(ràte)(tì) | *! | * | * | |
| c. (lng)(karà)(tetì) | | | *! | |

b. Grammar after exposure to odd-parity word

| Input: *lágkaràtetì* | DIR-R | HEAD-R | PARSE-σ | NONFIN |
|---|---|---|---|---|
| a. (langk)(ratè)ti | | | *! | |
| b. (lngka)(ràte)(tì) | *! | * | | * |
| c. ☞ (lng)(karà)(tetì) | | | | * |

certain candidates over others, in the absence of knowledge about which aspects of the current grammar are correct. The cue-based learner follows an ordered path and looks for patterns, not individual forms or sentences to match. By so doing, it is able to use information gained at earlier parts of the learning path to inform its progress later on the path. Crucially, the learner considers this information to be reliable, and in fact it is, unless the learner is being fooled. The various other learning algorithms we have reviewed all have different ways of moving the learner through the space of possible grammars, but there is never a guarantee that any move results in a better grammar, as opposed to an improved match to the current target form.

Tesar and Smolensky's distinction between overt and covert aspects of the target forms is subtly different from the approach taken here, and I call attention to it as a way of underscoring one of the main themes of this article. At the outset of the learning path sketched above, stress contours were taken to be overt, as were syllables and segments (but not feet, or the distinction between heavy and light syllables). Presumably, there are earlier stages of acquisition where even these phenomena are not overt to a learner: the segmentation of words into syllables is not entirely evident, nor are the acoustic cues that signal stress in any given language. Conversely, the distinction between light and heavy syllables is ''covert'' at early stages of acquisition but ''overt'' later, after the learner has fixed this distinction in the grammar. Thus, the categories ''overt'' and ''covert'' are also fluid, and they shift as the learner acquires more of the grammar.[34]

[34] This picture thus has something of a Piagetian flavor, in that subsequent stages are in some sense constructed out of earlier ones, but without the mystery of how the learner creates the new stage from the old; see the debate recorded in Piattelli-Palmarini 1980. In Piagetian terms, the account I am proposing remains stolidly in the ''preformationist'' camp.

## 6 Conclusion

To conclude, the thesis defended here is that an ordered cue-based learner of the type sketched in (2) offers the most promising approach to solving the fundamental problems of grammar acquisition set out in (1). On this approach, representations are gradually elaborated in the course of acquisition, guided by a set of ordered cues that become increasingly abstract and grammar-internal. I have argued that the other learning models reviewed above do not provide satisfactory solutions to one or both of the Credit Problem and the Epistemological Problem. In particular, they all treat as fixed and extensional (external to the learner) representations and relations that are properly intensional (internal to the learner) and constantly changing.

One other important characteristic distinguishes the cue-based learner from all the other proposals. In the other models, there is a sharp distinction between the learning algorithm and the grammar. Moreover, except at the most general level (whether it deals with parameters or violable constraints, for example), the learning algorithm is independent of the content of the grammar. Thus, once we have decided on a set of parameters/constraints, the operation of the learning algorithm proceeds automatically. We have seen, for example, that it makes no difference to the TLA what the content of a parameter is: the same chart serves for syntactic word order parameters as for parameters of metrical theory, or even for nonlinguistic parameters.

In the cue-based model, by contrast, the learning algorithm is not independent of the content of the parameters. The determination of the learning path is an empirical matter, which must be established with respect to each parameter. Having arrived at a learning path for metrical parameters, for example, we cannot simply slot syntactic parameters into their place to arrive at a learning algorithm for the acquisition of word order. Rather, we must establish anew what the cues and their ordering are for this domain. Thus, in the cue-based model the learning path is part of linguistic theory. It is an empirical issue what the correct relation between the grammar and the learning algorithm is. The hypothesis embodied by a cue-based learning model is that the relation is very close and that there is no general learning algorithm independent of the content of the grammar.

Finally, it should be observed that none of the models discussed here take adequate account of the actual course of development exhibited by children. The next step is to attempt to incorporate the results of research in this area.[35] These data show even more forcefully that the target input forms to the learner are moving targets, not given in advance of applying a learning algorithm. Rather, representations at all levels of the grammar are mental constructs, themselves the results of the acquisition of grammar.

## References

Avery, Peter. 1996. The representation of voicing contrasts. Doctoral dissertation, University of Toronto, Toronto, Ont.

Berwick, Robert C. 1985. *The acquisition of syntactic knowledge*. Cambridge, Mass.: MIT Press.

[35] See, for example, the dissertations by Fikkert (1994), Levelt (1994), and Nouveau (1995), which, taken together, paint a vivid picture of stages in the acquisition of Dutch phonology.

Berwick, Robert C., and Partha Niyogi. 1996. Learning from triggers. *Linguistic Inquiry* 27:605–622.

Chomsky, Noam. 1981. Principles and parameters in syntactic theory. In *Explanation in linguistics: The logical problem of language acquisition,* ed. Norbert Hornstein and David Lightfoot, 32–75. London: Longman.

Clark, Robin. 1989. On the relationship between the input data and parameter setting. In *Proceedings of NELS 19,* 48–62. GLSA, University of Massachusetts, Amherst.

Clark, Robin. 1990. *Papers on learnability and natural selection.* Technical Reports in Formal and Computational Linguistics, No. 1, Université de Genève.

Clark, Robin. 1992. The selection of syntactic knowledge. *Language Acquisition* 2:83–149.

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

Dinnsen, Daniel A. 1992. Variation in developing and fully developed phonetic inventories. In *Phonological development: Models, research, implications,* ed. Charles A. Ferguson, Lise Menn, and Carol Stoel-Gammon, 191–210. Timonium, Md.: York Press.

Dresher, Bezalel Elan. 1992. A learning model for a parametric theory in phonology. In *Formal grammar: Theory and implementation,* ed. Robert Levine, 290–317. New York: Oxford University Press.

Dresher, Bezalel Elan. 1993. Cues and parameters in phonology. In *CLS 27.* Part 1, *The General Session,* 119–140. Chicago Linguistic Society, University of Chicago, Chicago, Ill.

Dresher, Bezalel Elan. 1994. Acquiring stress systems. In *Language computations,* ed. Eric Sven Ristad, 71–92. Providence, R.I.: AMS.

Dresher, Bezalel Elan. 1996. Learnability and phonological theory. In *Current trends in phonology: Models and methods, vol. 1,* ed. Jacques Durand and Bernard Laks, 245–266. Salford: European Studies Research Institute, University of Salford.

Dresher, Bezalel Elan. Forthcoming. Child phonology, learnability, and phonological theory. In *Handbook of language acquisition,* ed. Tej K. Bhatia and William C. Ritchie. San Diego, Calif.: Academic Press.

Dresher, Bezalel Elan, and Harry van der Hulst. 1995. Head-dependent asymmetries in phonology. In *Leiden in last: HIL phonology papers I,* ed. Harry van der Hulst and Jeroen van de Weijer, 401–431. The Hague: Holland Academic Graphics.

Dresher, Bezalel Elan, and Harry van der Hulst. 1997. Head-dependent asymmetries in phonology: Complexity and visibility. Ms., University of Toronto, Toronto, Ont., and HIL/University of Leiden.

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

Fikkert, Paula. 1994. *On the acquisition of prosodic structure.* Dordrecht: ICG.

Fodor, Janet Dean. 1998. Unambiguous triggers. *Linguistic Inquiry* 29:1–36.

Fodor, Janet Dean. In press a. Learnability theory: Decoding trigger sentences. In *Linguistics, cognitive science, and childhood language disorders,* ed. Richard C. Schwartz. Hillsdale, N.J.: Lawrence Erlbaum.

Fodor, Janet Dean. In press b. Learnability theory: Triggers for parsing with. In *The development of second language grammars: A generative approach,* ed. Elaine C. Klein and Gita Martohardjono. Amsterdam: John Benjamins.

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

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

Gillis, Steven, and Gert Durieux. 1996. Data-driven approaches to phonological acquisition: An empirical test. In *Proceedings of the UBC International Conference on Phonological Acquisition,* ed. Barbara Bernhardt, John Gilbert, and David Ingram, 277–292. Somerville, Mass.: Cascadilla Press.

Gillis, Steven, Gert Durieux, and Walter Daelemans. 1995. A computational model of P&P: Dresher &

Kaye (1990) revisited. In *Approaches to parameter setting,* ed. Frank Wijnen and Maaike Verrips, 135–173. Vakgroep Algemene Taalwetenschap, Universiteit van Amsterdam.

Goedemans, Rob, Harry van der Hulst, and Ellis Visch, eds. 1996. *Stress patterns of the world.* Part 1, *Background.* The Hague: Holland Academic Graphics.

Halle, Morris, and G. N. Clements. 1983. *Problem book in phonology*. Cambridge, Mass.: MIT Press.

Halle, Morris, and William J. Idsardi. 1995. General properties of stress and metrical structure. In *The handbook of phonological theory,* ed. John Goldsmith, 403–443. Cambridge, Mass.: Blackwell.

Halle, Morris, and Jean-Roger Vergnaud. 1987. *An essay on stress.* Cambridge, Mass.: MIT Press.

Hayes, Bruce. 1981. A metrical theory of stress rules. Doctoral dissertation (1980), MIT, Cambridge, Mass. [Revised version distributed by Indiana University Linguistics Club, Bloomington, and published by Garland Press, New York, 1985.]

Hayes, Bruce. 1987. A revised parametric metrical theory. In *Proceedings of NELS 17,* 274–289. GLSA, University of Massachusetts, Amherst.

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

Hulst, Harry van der. 1994. Radical CV phonology: The locational gesture. In *UCL working papers 6,* 439–478. Department of Phonetics and Linguistics, University College London.

Idsardi, William J. 1992. The computation of prosody. Doctoral dissertation, MIT, Cambridge, Mass.

Ingram, David. 1989. *First language acquisition: Method, description, and explanation.* Cambridge: Cambridge University Press.

Jakobson, Roman. 1941/1968. *Child language, aphasia, and phonological universals.* The Hague: Mouton (1968). Translation by Allan R. Keiler of *Kindersprache, Aphasie, und allgemeine Lautgesetze.* Uppsala: Uppsala Universitets Årsskrift (1941).

Jakobson, Roman, and Morris Halle. 1956. *Fundamentals of language*. The Hague: Mouton.

Kager, René. 1993. Alternatives to the iambic-trochaic law. *Natural Language & Linguistic Theory* 11: 381–432.

Kapur, Shyam. 1994. Some applications of formal learning theory results to natural language acquisition. In *Syntactic theory and first language acquisition: Crosslinguistic perspectives.* Vol. 2, *Binding, dependencies, and learnability,* ed. Barbara Lust, Gabriella Hermon, and Jaklin Kornfilt, 491–508. Hillsdale, N.J.: Lawrence Erlbaum.

Kaye, Jonathan, Jean Lowenstamm, and Jean-Roger Vergnaud. 1985. The internal structure of phonological elements: A theory of charm and government. *Phonology Yearbook* 2:305–328.

Kemenade, Ans van. 1987. *Syntactic case and morphological case in the history of English.* Dordrecht: Foris.

Kroch, Anthony, and Ann Taylor. 1997. The syntax of verb movement in Middle English: Dialect variation and language contact. In *Parameters of morphosyntactic change,* ed. Ans van Kemenade and Nigel Vincent, 297–325. Cambridge: Cambridge University Press.

Levelt, Clara C. 1994. *On the acquisition of place*. Dordrecht: IGC.

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

Lightfoot, David. 1989. The child's trigger experience: Degree-0 learnability. *Behavioral and Brain Sciences* 12:321–375.

Lightfoot, David. 1991. *How to set parameters: Arguments from language change*. Cambridge, Mass.: MIT Press.

Lightfoot, David. 1997. Shifting triggers and diachronic reanalyses. In *Parameters of morphosyntactic change,* ed. Ans van Kemenade and Nigel Vincent, 253–272. Cambridge: Cambridge University Press.

Lightfoot, David. Forthcoming. Creoles and cues. In *Language creation and language change,* ed. Michel DeGraff. Cambridge, Mass.: MIT Press.

Macken, Marlys A. 1978. Permitted complexity in phonological development. *Lingua* 44:219–253.

Marcus, Mitchell P. 1980. *A theory of syntactic recognition for natural language*. Cambridge, Mass.: MIT Press.

McCarthy, John J., and Alan Prince. 1993. Generalized alignment. In *Yearbook of morphology 1993,* ed. Geert Booij and Jaap van Marle, 79–153. Dordrecht: Kluwer.

Nouveau, Dominique. 1994. *Language acquisition, metrical theory, and optimality*. Utrecht: Onderzoeksinstituut voor Taal en Spraak.

Nyberg, Eric H., 3rd. 1991a. A limited non-deterministic parameter-setting model. In *NELS 21,* 309–322. GLSA, University of Massachusetts, Amherst.

Nyberg, Eric H., 3rd. 1991b. A non-deterministic, success-driven model of parametric setting in language acquisition. Doctoral dissertation, Carnegie Mellon University, Pittsburgh, Pa.

Piattelli-Palmarini, Massimo, ed. 1980. *Language and learning: The debate between Jean Piaget and Noam Chomsky.* Cambridge, Mass.: Harvard University Press.

Prince, Alan. 1983. Relating to the grid. *Linguistic Inquiry* 14:19–100.

Prince, Alan. 1990. Quantitative consequences of rhythmic organization. In *CLS 26.* Vol. 2, *Parasession on the Syllable in Phonetics and Phonology,* 355–398. Chicago Linguistic Society, University of Chicago, Chicago, Ill.

Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical report RuCCS-TR2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, N.J.

Rice, Keren D. 1992. On deriving sonority: A structural account of sonority relationships. *Phonology* 9: 61–99.

Rice, Keren D., and Peter Avery. 1995. Variability in a deterministic model of language acquisition: A theory of segmental elaboration. In *Phonological acquisition and phonological theory,* ed. John Archibald, 23–42. Hillsdale, N.J.: Lawrence Erlbaum.

Tesar, Bruce. 1997. An iterative strategy for learning metrical stress in Optimality Theory. In *Proceedings of the 21st Annual Boston University Conference on Language Development,* ed. Elizabeth Hughes, Mary Hughes, and Annabel Greenhill, 615–626. Somerville, Mass.: Cascadilla Press.

Tesar, Bruce. 1998. An iterative strategy for language learning. *Lingua* 104:131–145.

Tesar, Bruce, and Paul Smolensky. 1993. The learnability of Optimality Theory: An algorithm and some basic complexity results. Technical report CU-CS-678-93, Department of Computer Science, University of Colorado, Boulder.

Tesar, Bruce, and Paul Smolensky. 1996. Learnability in Optimality Theory (long version). Technical report JHU-CogSci-96-3, Department of Cognitive Science, Johns Hopkins University, Baltimore, Md.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Tryon, Darrell T. 1970. *An introduction to Maranungku*. Canberra: Australian National University.

Turkel, William J. 1996. Smoothness in a parametric subspace. Ms., University of British Columbia, Vancouver.

Tyler, Stephen A. 1969. *Koya: An outline grammar - Gomma dialect.* Berkeley and Los Angeles: University of California Press.

Waterson, Natalie. 1978. Growth of complexity in phonological development. In *The development of communication,* ed. Natalie Waterson and Catherine E. Snow, 415–442. New York: Wiley. [Revised version in Waterson 1987, 88–107.]

Waterson, Natalie. 1987. *Prosodic Phonology: The theory and its application to language acquisition and speech processing.* Newcastle upon Tyne: Grevatt & Grevatt.

*Department of Linguistics*
*University of Toronto*
*Toronto, Ontario*
*Canada M5S 3H1*

*dresher@chass.utoronto.ca*